Contrôle du FDR et imputation de valeurs manquantes pour l'analyse de données de protéomiques par spectrométrie de masse

Lucas Etourneau

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : ISCE - Ingénierie pour la Santé la Cognition et l'Environnement
Spécialité : MBS - Modèles, méthodes et algorithmes en biologie, santé et environnement
Unité de recherche : BGE - Laboratoire Biosciences et bioingénierie pour la Santé

Contrôle du FDR et imputation de valeurs manquantes pour l'analyse de données de protéomiques par spectrométrie de masse

FDR control and missing value imputation for the analysis of mass spectrometry-based proteomics data

Présentée par :

Lucas ETOURNEAU

Direction de thèse :

Thomas BURGER DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES	Directeur de thèse
Nelle VAROQUAUX CHARGEE DE RECHERCHE, CNRS DELEGATION ALPES	Co-encadrant de thèse

Rapporteurs :

JULIE JOSSE
SENIOR SCIENTIST, ANTENNE INRIA UNIVERSITE DE MONTPELLIER
NATALIYA SOKOLOVSKA
PROFESSEURE DES UNIVERSITES, SORBONNE UNIVERSITE

Thèse soutenue publiquement le **24 janvier 2024**, devant le jury composé de :

THOMAS BURGER DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES	Directeur de thèse
JULIE JOSSE SENIOR SCIENTIST, ANTENNE INRIA UNIVERSITE DE MONTPELLIER	Rapporteure
NATALIYA SOKOLOVSKA PROFESSEURE DES UNIVERSITES, SORBONNE UNIVERSITE	Rapporteure
ADELINE LECLERCQ SAMSON PROFESSEURE DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Présidente
GUILLAUME FERTIN PROFESSEUR DES UNIVERSITES, UNIVERSITE DE NANTES	Examinateur
QUENTIN GIAI-GIANETTO INGENIEUR DE RECHERCHE, INSTITUT PASTEUR	Examinateur

Invités :

NELLE VAROQUAUX
CHARGEE DE RECHERCHE, CNRS DELEGATION ALPES

À mamie Georgette, papi René, papi Antoine et Agnès

"Le mystère de la vie n'est pas un problème à résoudre, c'est une réalité dont il faut faire l'expérience."

J. J. Van Der Leeuw, dans *La Conquête de L'Illusion* (1928)

Remerciements

Je tiens premier lieu à remercier Mme Julie Josse et Mme Nataliya Sokolovska pour avoir accepté de rapporter cette thèse, ainsi que Mme Adeline Leclercq Samson, Mr Quentin Giai-Gianetto et Mr Guillaume Fertin d'avoir accepté d'évaluer cette thèse.

Mes seconds remerciements vont à mes encadrants, Thomas Burger et Nelle Varoquaux, qui d'une part, ont accepté de me confier cette thèse, et d'autre part, m'ont accompagné sans relâche durant ces trois années. Leur implication dans mon projet, leurs innombrables conseils et leur bienveillance m'ont permis de progresser sur de nombreux aspects de la recherche (interdisciplinaire) et m'ont aidé à garder une motivation presque intacte tout du long. Je leur dois amplement tout ce travail accompli, et leur suis sincèrement reconnaissant pour tout cela.

Je tiens ensuite à remercier les membres de mon comité de suivi, Clovis Galiez et Julien Chiquet. La première année fût assez rude pour moi, en partie du fait de la pandémie. Ils ont non seulement su me remotiver en une demi-journée, mais m'ont donné de précieux conseils lors nos rares entrevues et je leur en remercie.

Je remercie ensuite mes plus proches collaborateurs durant cette thèse, à commencer par Laura Fancello. Ses cours particuliers en biologie moléculaire ont grandement contribué à combler mes lacunes, et son implication dans mon projet a été d'une très grande aide. Tout le travail qu'elle a laissé derrière était d'une propreté impeccable et d'une clarté limpide, et je la remercie sincèrement. Ensuite, Samuel Wieczorek m'a considérablement aidé sur l'implémentation du package, ce qui m'a permis de me concentrer sur la rédaction, si cruciale dans ces derniers mois de thèse, merci à lui. Enfin, j'ai eu la chance (et l'honneur!) que David Pérez me propose de collaborer avec lui sur son sujet de thèse. C'était réellement un plaisir de pouvoir échanger, réfléchir, aboutir à des solutions et des résultats ensemble, et avec un copain qui plus est. Je suis fier de ce qu'on a fait.

D'autres membre d'EDyP ont également apporté leur pierre non négligeable à l'édifice. Je pense notamment à Véronique Dupierris qui a pris du temps et s'est arrachée quelques cheveux pour me fournir des ressources de calculs importantes, Julia Novion Ducassou et Anne-Marie Hesse qui m'ont donné de précieux cours en MS (Anne-Marie m'ayant aussi fourni des datasets tout propres), Christophe Bruley et Yohann Couté qui m'ont accompagné, posé de nombreuses questions pertinentes, ainsi qu'un regard affuté sur ma recherche.

Du côté du TIMC, je tiens à remercier Sophie Abby, Ivan Junier et Antoine Frenoy qui m'ont tous écouté parler moulte fois, bien que leur thématique de recherche soit différente. Ils m'ont fourni de précieux conseils, à chaque fois accompagnés de remarques pertinentes.

Pour conclure sur ces remerciements concernant mes travaux, je souhaiterais mentionner mes quelques discussions avec Quentin Giai Gianetto qui m'ont beaucoup apporté. Son retour d'expérience, et le partage de sa vision sur certains problèmes en protéomique statistique m'ont grandement inspiré, et je lui en suis reconnaissant.

Place enfin, non sans émotions, aux remerciements d'ordre plus personnels.

Tout d'abord, merci à l'ensemble de l'équipe d'EDyP pour le cadre de travail serein et bienveillant qu'elle a su créer. Nos pauses dej/café/autres ont été des moments de décompression intense, accompagné d'éclat de rire de toute part. À toute l'équipe "midi H3", à qui j'ai si souvent manqué de respect en me présentant à 12h03, et avec qui j'ai dépensé une fortune en blonde IPA, c'est-à-dire Louis, Vaitson, David, Julia, Hassan, Victoria, Wioletta, Thijs, Simon, Romain, Nicolas

et Naomi, votre bienveillance, gentillesse et votre humour, presque toujours au niveau, restera dans ma mémoire. Louis, avec qui j'ai passé plus de temps à l'extérieur du CEA qu'à l'intérieur (on me rétorquera peut-être que ce n'était pas très compliqué) pour vivre des aventures extraordinaires, merci à toi du fond du cœur pour ta simple présence.

Le même remerciement va à l'ensemble de l'équipe TrEE, et en particulier aux membres du groupe bioinfo. Sophie-Carole, Morgane, Margaux et Katayoun, vous avez créé une athmosphère décontractée et bienveillante dans l'équipe et je vous souhaite le meilleur pour la fin de vos thèses et post-doc. Votre humour n'a absolument rien à envier à mes copains et copines du CEA, et j'ai vraiment passé de très bons moments à vos côtés. Sophal, it was a real pleasure to meet you, share some deep thoughts and sport time together. I wish we could have spent more time together, but it's only partie remise, comme on dit ici.

Je n'oublie pas tous mes copains et copines de Grenoble, Paris et d'ailleurs qui m'ont fait sortir de ma petite bulle de thèse, à travers des voyages, festivals et autres, et que je compte bien continuer à voir. Eloi, c'est toujours un vrai bonheur de discuter avec toi de sciences, de tout et de n'importe quoi, ainsi que d'annihiler ensemble des hordes de zombies.

Merci également à ma famille, même si cela dépasse largement ces trois années de thèse. Tout d'abord mes parents, à jamais fidèles au poste, par leur soutien, leur présence et leur amour incommensurable, merci. Ma sœur, de sang et d'esprit, ma meilleure amie pour la vie comme dirait un collégien de 12 ans, sauf que cette fois c'est vrai, merci. Et puis ma grand-mère qui, bien qu'avare de questions, n'en pense jamais moins, a toujours veillé à ce que je ne manque pas de nourriture. Je lui dois beaucoup.

Enfin, je terminerai par le king, le boss, le chef, el famoso, également dénommé la Gauffre (oui oui avec deux f), je parle bien ici d'Aurélien, mon coloc/copain/complice/community manager (bref tout ce qui commence par co-). La thèse aurait probablement une autre saveur sans avoir vécu au quotidien avec lui, et ça a été un honneur de servir la science à ses côtés. Tout est parti d'une discussion en Bretagne qui peut se résumer à : "- Je viens d'accepter une thèse à Grenoble, tu veux rester à là-bas et habiter avec moi pendant 3 ans ? - Oui". Quelle chance j'ai eu ce jour-là. Le plaisir était intense, les souvenir resteront immenses. Aurélien, merci pour tout.

Contents

General introduction

Proteins are the building blocks of life. They serve multiple roles, from participating in cellular mechanisms to forming the structure of the cell. Their expression is dynamic, varying not just across different cell types and tissues but also temporally. Molecular biology is committed to studying the complex processes at stake in gene expression, and to do so, it notably relies on proteomics, which purpose is to identify and quantify the complex mixture of proteins found in biological samples, at large-scale.

To do so, high performance liquid chromatography coupled to tandem mass spectrometry (a.k.a. LC-MS/MS) is the most widespread method. The work presented here revolves around it, and more specifically around the so-called label-free bottom-up LC-MS/MS analyses: This type of analysis notably involves the characterization of protein fragments called peptides, in each sample independently. This technology generates massive amounts of data, from which the extraction of biological knowledge requires specific statistical processing. Briefly, peptides must be identified (by comparing the MS signals to the peptides' theoretical signature), and quantified (thanks to the MS signal intensity). This leads to a data summary, usually of tabular form, listing peptides and their abundance, which is of prime importance to discover biomarkers of specific phenotypes.

Several challenges are inherent in this data processing pipeline. First, two important steps often contribute to false positives: peptide identification and biomarker identification. The former is challenging due to the numerous comparisons between noisy experimental spectra and their theoretical counterparts. The latter is not easier, given the multitude of candidate proteins that are tested for biomarker discovery. These problems intersect at the need for false discovery rate (FDR) control, which aims at determining appropriate thresholds for retaining peptide identifications or potential biomarkers of sufficiently high confidence as to minimize false positives (*i.e.*, incorrect identifications or spurious biomarkers). Another major issue relates to missing values (MVs), which are abundant and cannot be ignored. Unfortunately, current imputation methods do not adequately address this challenge, notably because of the complex and manifold origin of these MVs.

This thesis presents contributions to the overcoming of these issues of FDR control and of missing value imputation. They are hereafter presented as follows:

- **Chapter I** introduces the basic principles of molecular biology and of LC-MS/MS proteomics. It more precisely defines the biostatistical issues tackled in this manuscript, the theoretical foundations necessary to solve them and the current solutions used in LC-MS/MS proteomics.

- **Chapter II** zooms in on the FDR control. It presents three publications, two of which revolve around a general FDR control framework referred to as *knockoff filters*, that we put in perspective regarding proteomic applications [Etourneau22, Etourneau23]. The last one extends biomarker selection (under FDR control constraints) to their combination to improve the diagnosis of a liver disease.

- **Chapter III** introduces a novel imputation algorithm specifically tailored for proteomic data, which can also incorporate transcriptomic quantitative information.

- We conclude with a summary of our findings, perspectives based on our presented works, and more general thoughts upon LC-MS/MS proteomics and missing value imputation.

1 Statistical proteomics background

We introduce here the main concepts of statistical proteomics, from the biological, chemical, computational, and mathematical point of views, as well as some important related challenges that are addressed in this thesis.

Sommaire

1.1 Applicative Context

1.1.1 The central dogma of biology

Proposed by F. Crick in 1957 [Cobb17], the central dogma of molecular biology is a powerful concept to describe the flow of genetic information at molecular level inside a biological system. We can sum it up as following: a *gene* is a piece of a long sequence of *nucleotides* (4 biomolecules pictured as A, C, G, T), called DNA (desoxyribonucleic acid). This sequence is transcribed into another sequence of nucleic acids, called *transcript* or mRNA (messenger ribonucleic acid). Finally, the transcript is translated into a coiled sequence of *amino acids* (AA), called a *protein*. Proteins are often referred as the "building blocks of life" as they are involved in most of the tasks of the cell life, through various functions [Alberts17]: they constitute the cell membrane, act as catalysts, receptors, switches, tiny motors or pumps, etc.

The function of a protein is largely determined by its 3D structure, which is itself largely determined by its amino acid sequence [Hinz10], and hence the information encoded in the transcript and gene sequences. Yet, relying only on the central dogma is limiting to understand the whole molecular dynamics resulting from genes. For example, while the central dogma ensures that information is transmitted from genes to proteins through transcripts, it does not provide quantitative insight on the expression of genes (*i.e.*, which quantity of the resulting proteins are produced). For a given gene inside a cell, the number of mRNA copies may vary dramatically over time, and the relation between the number of mRNA copies and the number of resulting protein copies depends on many biochemical and environmental factors [Liu16]. For example, small RNAs (short non-coding RNA sequences) can hinder the expression of more than 60% of protein-coding genes, either by disabling mRNA translation by docking on them, or the gene transcription by altering chromatin configuration (the support of DNA) [Stuwe14]. Therefore, a high transcription level at a given time in a cell does not necessary implies a large protein amount, and vice-versa. Also, a same gene can be transcribed into different mRNAs, then referred to as *isoforms*, because of the alternative splicing mechanism [Marasco22]. It refers to the fact that some parts of the initial mRNA are not kept for translation, and the remaining parts can be reassembled in various ways. Additionally, post-translational modifications (a.k.a. PTMs –modifications occurring after translation, which can change a function of a protein or affect its biological activity [Uversky13]) can occur and this additional information cannot be derived from transcript analysis. Alternative splicing and PTMs are typical of the exploding combinatory leading to proteins configurations, which drastically complicates the study of the gene expression products.

1.1.2 The omic paradigm

Since 1953 and the discovery of the structure of DNA by J. Watson and F. Crick, research in genetics have essentially resulted in an analytical approach, by isolating a given gene, its different interactions and forms, and depicting an exhaustive description of it [Lay06]. However, in the 90's, the development of sequencing technologies leveraging increasingly powerful computational resources has enabled a new complementary approach, privileging a global view of gene information. This approach was coined genomic, as the discipline of the genome, the kingdom of genes.

From that point on, molecular biology has thrived into multiple "omic" fields (an almost exhaustive list of those that have emerged thanks to the development of various high-throughput and cheapest instruments is given on this page [Wikipedia23]). To date, any field coined *thingy*-omics refers to the study of the *thingy* type of biomolecule at *large-scale*. By large-scale, one means an attempt to characterize (identify and quantify) the kingdom of *thingies* (the *thingies*-ome) within a sample as exhaustively as possible. Then, with bioinformatics and statistical tools, one then extracts knowledge from these large-scale (a.k.a. *high-throughput*) *thingy*-omics data.

Accordingly, the transcriptome and proteome have respectively been defined as the set of all transcripts and of all proteins contained inside an organism, tissue, or cell. Their associated

discipline, the transcriptomics [Lowe17] and the proteomics [Anderson98] focus on gene expression, as to give a snapshot of the current biological state of an organism, tissue or cell (the *phenotype*), which explains their particular interest for biology and medical sciences. Indeed, unlike the genome, which remains relatively stable across time and different environmental conditions, the phenotype is dynamic. As another example, the genome of the caterpillar and resulting butterfly is the same, but their phenotype differs drastically. However, genomics remain mandatory to address hereditary questions. More generally, each of the omics sheds a specific light on the molecular interactions in an organism, thus offering complementary biological information. For example, metabolomics (metabolites are product of chemical reactions inside an organism) covers a set of biomolecules that are not mentioned in the central dogma (even though these chemical reactions often imply proteins). Hence the direct measurement of metabolome provides a better understanding of the physiology and mechanistic of cells, which would be hardly achievable otherwise [Villate21]. Also, some omics can help to fill the gap of others. As it will appear latter, proteomics and transcriptomics do not provide the same access to the phenotype; and to identify proteins in a proteomic analysis, one classically relies on sequence databases derived from the genomic studies of the specie of interest.

1.1.3 Measuring gene expression by transcriptomics and proteomics

Proteomics and transcriptomics both aim at measuring gene expression at large scale. We give here an overview of the two approaches, the current technologies used and comparative elements.

Transcriptomic analysis today relies on RNA-Seq technology (often considered among Next Generation Sequencing methods [Behjati13]), which sequences and counts the transcript fragment (referred to as *reads*) in the sample. It is a powerful technology as it requires no prior knowledge on the sequences, and it almost exhaustively covers the transcriptome. Yet, the processing of read counts is still challenging. For instance, whether zero count for a given read should be interpreted as a zero measurement or an absence of measure, is still controversial [Silverman20]. A fast-emerging field is single-cell transcriptomics (referred to as scRNA-Seq). As opposed to bulk-level analysis, scRNA-Seq enables the resolution of transcriptome at cell level, and for thousands of them. Hence, scRNA-Seq experiments can capture the cell variability over a tissue, which has particular interest in oncology for example [Kanter15].

The field of proteomics aims at characterizing and quantifying the proteome [Anderson98]. For the reasons mentioned above, it brings a finer level of understanding of gene expression than the sole study of transcriptome. An important application lies in finding biomolecular differences, *i.e.*, biomarkers, between different biological conditions (*e.g.*, healthy and diseased patients), which can lead to more precise diagnosis in clinical context. More generally, proteomics analysis helps biologists to understand differences between different cells, tissues, wild-type and mutated species etc. The domain has gained in popularity over the last decades, and to date, many biomedical research teams rely on proteomics for their investigations. Although one lab may not be representative, this is illustrated by the activity of the proteomic service platform partnering the lab where this thesis was prepared (EDyP lab): it contributed to 50 publications in 2012 and to 130 publications in 2022.

High-throughput proteomics mostly relies on mass-spectrometry coupled with liquid chromatography (LC-MS/MS). Other methods such as enzyme-linked immunosorbent assay (ELISA, [Anderson98]), relying on antibody linking, has long been used to precisely quantify proteins in a sample, however these are mostly targeted methods, *i.e.*, they can only deal with a limited number of different proteins. On the opposite, LC-MS/MS enables a wide covering of the proteome, and unlike RNA-Seq, offers a direct view to the phenotype (rather than a gene transcription-based proxy). However, LC-MS/MS still has some limitations. First, LC-MS/MS experiments mostly rely on databases of known proteins and PTMs, preventing the discovery of new proteoforms (*i.e.*, the different forms of a protein produced from the genome with a variety of sequence variations, splice isoforms, and PTMs). Alternative methods such as *de novo* sequencing in LC-MS/MS may cope with this, but current methods lack accuracy for a stand-alone use [Muth18]. Secondly, the

cost of analysis of a single sample in LC-MS/MS experiments oscillates between 700€ and 1100€ whereas one RNA-Seq analysis costs between 100€ and 200€. Recent methods in proteomics allow sample multiplexing (notably TMT technology [Thompson03]), so that several samples can be analyzed in the same LC-MS/MS experiment. However, the effective gain of cost becomes significant for a high number of samples only, and the quantifications obtained are often prone to distortions [Pappireddi19]. Finally, and probably most importantly, the coverage of LC-MS/MS, in terms of number of genes expressed, is limited compared to the RNA-Seq technology, mainly because of the sensitivity of the instrument (more detailed explanations in the next section).

Although LC-MS/MS based proteomics has up to date some limitations regarding the RNA-Seq approach, its potential to directly unveil gene expression at an unprecedented level justifies the efforts made in methodological development, from the analytical, bioinformatics, and statistics standpoints.

1.1.4 Large-scale proteomics by label-free LC-MS/MS

This section summarizes the main steps of a standard bottom-up label-free LC-MS/MS proteomic experiment (see Figure 1.1), also referred to as "discovery proteomics," as well as the processing of the resulting data, both according to the typical workflows at use in EDyP lab (many variants or alternative approaches described in the literature are thus ignored).

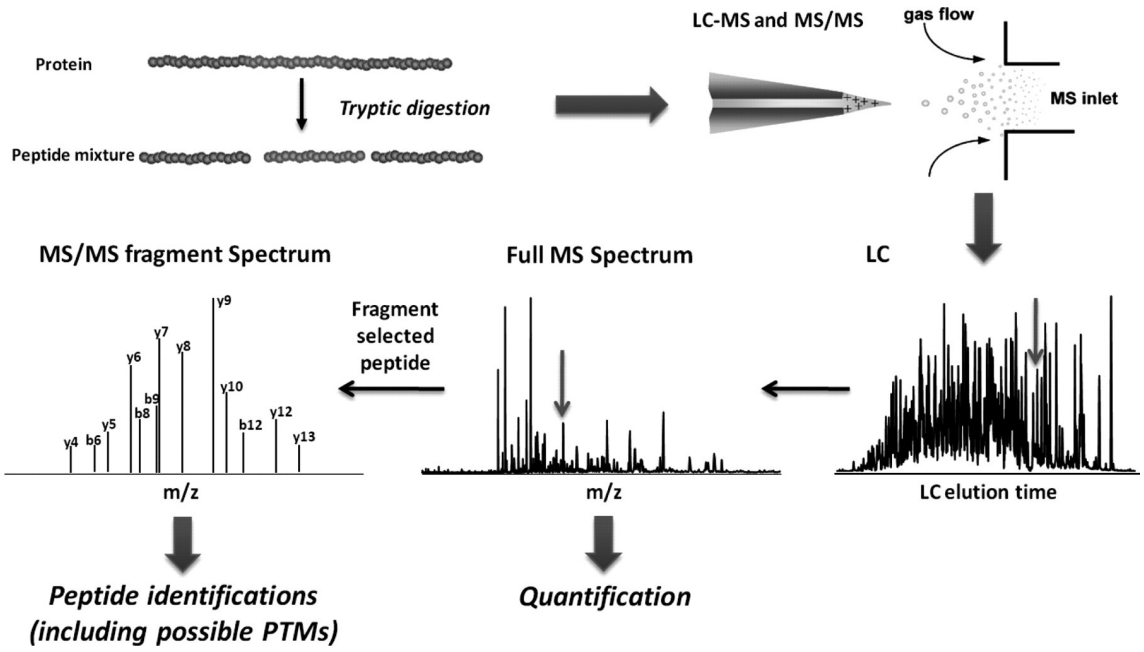

Figure 1.1: The bottom-up proteomic workflow, in data-dependent acquisition mode (from F. Xie et al. [Xie11]). 1. Proteins are digested by specific enzymes into peptides. 2. The peptide mixture is separated by LC and ionized before entering the mass spectrometer. 3. Full MS spectrum is acquired for the peptides that are eluting from the LC column at any given time. 4. In a data-dependent acquisition setting, one of the most intensive ion species (i.e., peptides) is then isolated and fragmented to obtain the MS pattern of its fragments, i.e., MS/MS spectrum). 5. The peptide sequence can be deduced from the MS/MS spectrum.

Biochemical and analytical treatment

First, proteins are extracted from a biological sample, and then digested by a specific enzyme (which cleaves the binding between specific patterns of amino acids), resulting in a complex mixture of *peptides* (*i.e.*, fragments of proteins, *i.e.*, shorter sequences of amino acids). This

explains the term "bottom-up" proteomics, as, starting from now in the process, one only deals with peptides instead of entire proteins. The reason is that peptides are easier to analyze by mass spectrometry than intact proteins [Tsiatsiani15, Gillet16]. Recovering protein-level information from peptide-level measurements (a step referred to as peptide-to-protein aggregation) is thus an essential computational step that is discussed later. To improve the sensitivity and the proteome coverage, the peptides are separated by liquid chromatography (LC) before entering the mass spectrometer (MS) [Niessen06]. The mass spectrometer then ionizes and analyzes the peptides on the fly as they elute from the LC, and, at each time stamp, produces an MS (or MS1, or Full MS) spectrum. Broadly, an MS spectrum is a plot of the intensity values measured as a function of the mass to charge ratio (denoted by m/z) of the measured ionized peptides (a.k.a. *precursor ions* or *precursors*). Yet, these spectra are not sufficient to identify the peptides analyzed for a simple reason: several peptides have exactly or almost the same mass, so that they would produce a peak at the same m/z point. To differentiate between them, the mass spectrometer isolates peptides in a chosen m/z region, fragments them and analyzes these fragments. This yields a second type of MS spectra, called MS/MS (or MS2, or fragmentation spectra), which contains the signature of the amino acid sequence, enabling identification of the peptide. Several m/z regions are analyzed, resulting in several MS/MS spectra per MS spectrum. This process of one MS followed by MS/MS acquisitions forms a cycle that is iterated during the whole elution time of the LC.

Peptide identification

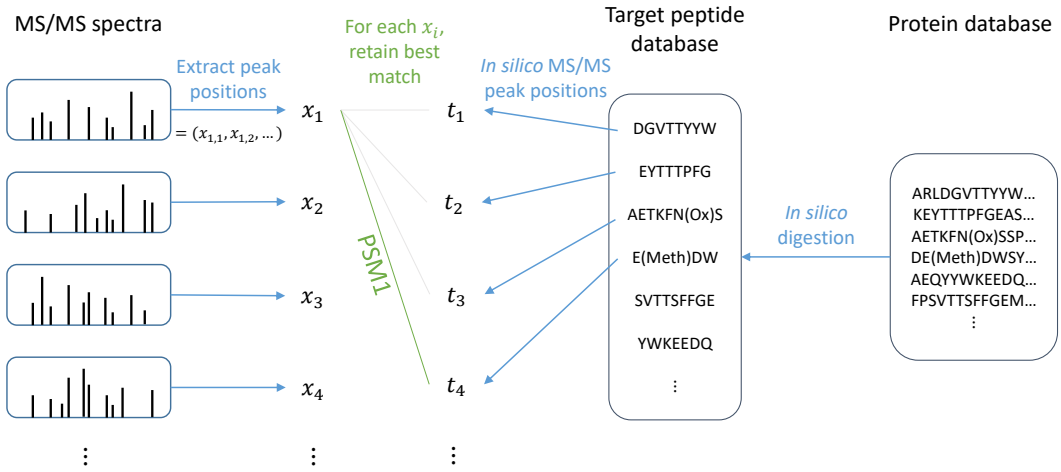

Figure 1.2: Illustration of peptide identification from MS/MS spectra in bottom-up label-free LC-MS/MS experiments.

There are two acquisition modes for MS/MS spectra: either data-dependent (DDA) or data-independent (DIA). In DDA mode, the different m/z regions selected for MS/MS correspond to the N most intense peaks in the MS1 spectrum (as suggested in Figure 1.1), where N is user-defined [Stahl96]. In DIA, the MS1 spectrum is decomposed in several m/z windows of same sizes (regardless of the spectrum content), which are then analyzed separately in MS/MS [Vidova17, Doerr14]. At first look, the DIA approach appears to be ideal, as it enables to cover the complete m/z range. Compared to DDA, in which only a few peaks are analyzed, it looks much more exhaustive. However, in DIA, each MS/MS spectrum contains a complex and multiplexed signal, as it depicts the fragmentation patterns of several peptides (which peaks are entangled). Thus, the identification of peptides in MS/MS requires a particular processing, often specific to the type of sample analyzed, and with variable performances regarding the tool used [Fröhlich22]. Conversely, in DDA, one MS/MS spectrum corresponds to a single peptide (unless two have the same mass and elute at same time, which occurs only occasionally), making the peptide identification more straightforward.

Even if some datasets produced in DIA have been used to benchmark our contributions, the rest of this chapter focuses on the DDA mode, as it is simpler and sufficient to set the basic principles that govern peptide identification.

Assuming each MS/MS spectra pertain to a single precursor ion (thus following DDA principles), it is searched against a reference (or target) database of peptide sequences, in the following manner (depicted in Figure 1.2). First, a database of protein sequences related to the species of interest is extracted from a large publicly available protein database (*e.g.*, RefSeq [Zahn-Zabal20], UniProt [Bateman21], or neXtProt [O'Leary16]). Note that these large databases heavily rely on the discovery of coding sequences in the genome, sequencing of variants and experts' annotations, which shows how advances in genomics and transcriptomics expand proteomic coverage. Then, *in silico* digestion and fragmentation of proteins is performed to obtain a list of target peptides, with their associated theoretical MS/MS peak positions. Finally, the list of peak positions of the experimental spectrum is compared with those of each item of the target database, and the best match (the notion of "best" being based on a tool-specific score) defines a peptide-spectrum match (PSM).

The whole process is performed automatically by search engines such as Mascot [Perkins99], Andromeda [Cox11], X!Tandem [Craig04], OMSSA [Geer04] and many others [Verheggen16]. As the number of pairwise comparisons to obtain PSMs is huge, we can expect that many of them occur by chance (the spectrum randomly matches the theoretical peaks of an unrelated sequence from the database). To cope with this, practitioners need to set a score threshold below which PSMs are assumed to depict a match of too poor quality to be relied on. Ideally, this threshold should guarantee that the probability of a random match is controlled and kept under a certain acceptable risk value. The computation of this score is a major issue of statistical proteomics and will be more extensively discussed in subsection 1.2.2.

Peptide quantification

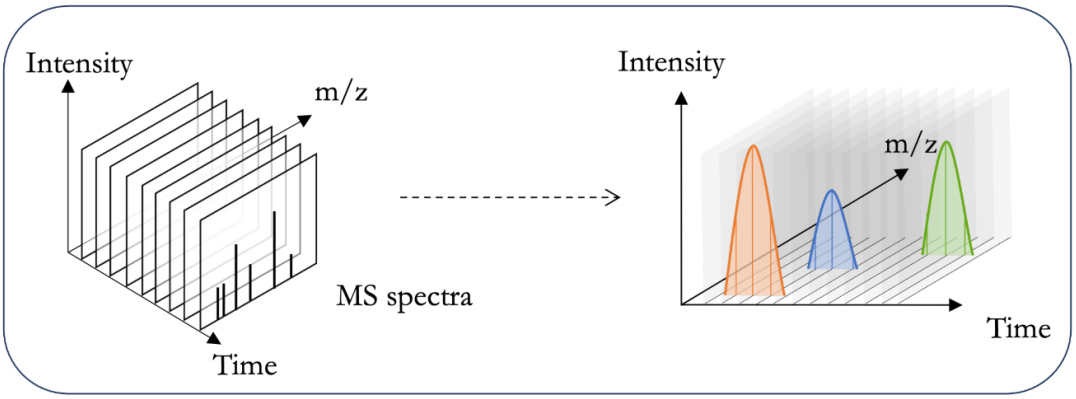

Figure 1.3: Illustration of the XIC peptide quantification by integrating the MS1 signal over elution time (from M. Chion and J. Bons [Chion21]).

The most classical approach to quantify peptides is to integrate the MS1 signal over the elution time, in the specific m/z window where each precursor has been identified (see Figure 1.3). Hence, peptides at a specific m/z value can be identified at few time points, but it does not ensure they can be quantified (the signal may be too discontinuous). This type of quantization is referred to as Extracted-Ion Chromatogram (XIC) [Bantscheff12].

The interpretation of peptide abundances constrains the downstream analysis of proteomics data resulting from LC-MS/MS experiments. The reason is the height of a peak in an MS1 spectrum does not only relate to the peptide quantity, but also to its chemical properties (which notably influence how it ionizes). Therefore, different XIC values cannot be compared unless they pertain

to the very same peptide sequence; quantitative comparisons only make sense in a relative way, for a same peptide across different samples. As a result, XIC quantification requires cautiousness: to compare XIC values across samples, the molecular composition of the samples should be similar. This is why some quality control steps are then required before computing XIC when quantification results are aggregated over different samples [Bateman14].

Of note, it is possible to mark proteins or peptides using isotopic standards as to derive absolute quantitation methods, but such *labelled-based* approaches are more cumbersome and do not easily scale-up to broad proteome analysis, so that this work has only focused on data produced with *label-free* yet relative quantitation approaches.

Filtering and normalization

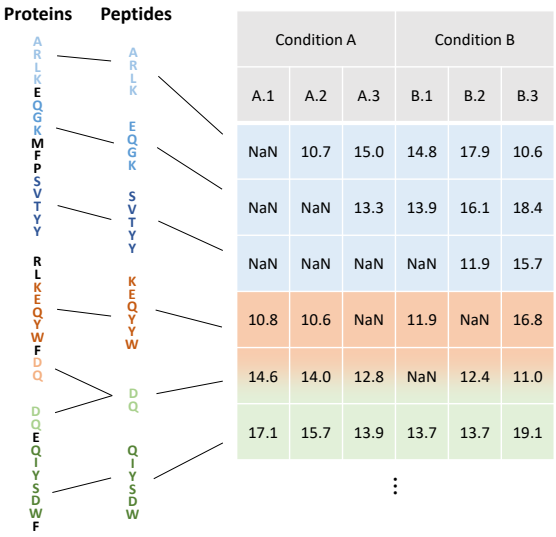

Figure 1.4: Illustration of the peptide abundance table obtained for a bottom-up label-free LC-MS/MS experiment.

At this step, the quantification results of different samples can be aggregated in a single table (illustrated on Figure 1.4), with peptides in rows, samples in columns, and an abundance value in each cell (note that in practice we deal with log-abundances to compress the expression scale, as well as to have broadly normally-distributed expression values in samples). As samples are analyzed independently, the non-overlapping identification and/or quantifications across samples induces missing values (MVs) in the table. Hence, some peptides with too many MVs are often filtered out from the table.

Then, intensities of samples may be normalized to correct unwanted variability between samples (*e.g.*, samples preparation, mass spectrometer calibration shift, etc.). Various normalization methods are used in proteomics: mean, median, or quantile alignment, as well as some other more specific ones developed for microarray gene expression data (Variance Stabilizing Normalization-VSN [Huber03], Removal of Unwanted Variation-RUV [Jacob16]) or metabolomic data (RUV for metabolomics [Livera15]). In any case, the choice of the method, and how it applies to the samples according to the experimental design is project specific.

Imputation of missing values

The peptide quantification table obtained contains an important amount of missing values. For example, it is not rare to have between 20 and 40% of MVs [Liu21, Webb-Robertson15], whereas

in many datasets, more than 50% of peptides have at least one MV. Yet, reader should keep in mind the MVs rate strongly varies according to the nature of the experiment, as MVs have many sources.

The first one is the non-overlapping of peptide identifications across samples. Hence, a dataset produced from the analysis of different types of tissues should have much more MVs than if it involves related samples: First, because the proteins may indeed differ between the tissues; And second, because the mass spectrometer has a certain sensitivity, such that abundant proteins identified in a tissue may not be so in another if their concentration is too low.

This leads us to the second source of MVs: the mass spectrometer has a dynamic range of view. It is comparable to a human pupil that dilates itself depending on the overall light intensity. Hence, the lower limit of detection of the instrument varies across elution time and from sample to sample, depending on the overall quantity of biological material currently analyzed. This results in a stochastic left-censoring of the abundances data, which is particularly difficult to account for.

Finally, a large proportion of MVs is expected to occur randomly in the dataset, due to the complexity of the biological samples and to the experimental procedure. For example, some peptides may not be cleaved as expected during digestion step, such that we cannot detect them properly. Without pretending to be exhaustive, we can give many other examples: misidentification, bad ionization, weak response in mass spectrometer, or scarce MS signal over elution time [Lazar16]. In any case, proteomists must keep in mind that there is no reliable way to determine the underlying cause of a given MVs, let it be because of low abundance or because of an issue in the experimental pipeline.

Several options are available to handle missing values and will be discussed in detail thereafter. Among them, imputation of missing values, *i.e.*, their replacement by a plausible value, is the approach chosen in EDyP Lab analysis pipeline.

Protein inference and aggregation of abundances

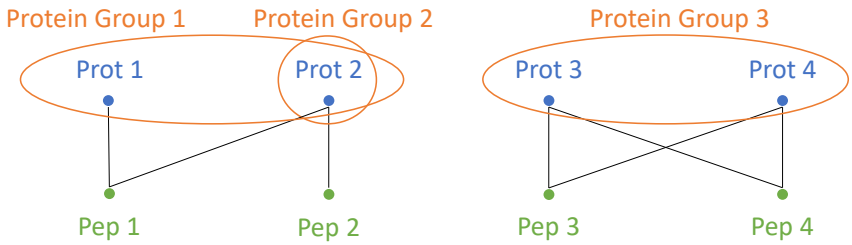

Figure 1.5: Illustration of protein groups. Prot 2 has a specific peptide (Pep 2) and thus defines its own protein group. Prot 1 only has one peptide, shared with Prot 2, and is thus grouped with Prot 2 in another protein group. Prot 3 and 4 are grouped as they only have peptides shared between them.

So far, the analytical and statistical steps have been conducted at peptide-level. However, proteins are the molecules of interest for measuring gene expression levels. For this reason, most of proteomic pipelines include a peptide-to-protein aggregation step. This requires first to identify the proteins represented by the peptide-level abundance table. When an identified peptide sequence is specific to a unique protein in the reference database, it simply proves the existence of the protein. However, an identified peptide sequence can also be shared between several proteins. Hence, if a protein does not have any specific peptide identified, it must be grouped with other proteins they share peptides with, leading to a so-called *protein group* (as its different constituting proteins cannot be discriminated using MS evidence), as illustrated on Figure 1.5. Second, once the proteins or protein groups are defined, one needs to endow them with an abundance value. This is usually done by averaging or summing (for each sample) the abundances of the peptides associated to the protein or the protein group [Blein-Nicolas16], although more sophisticated methods have been

developed [Zhang15, Bubis17]. A major difficulty is that, by the nature of abundance measurements (see section 1.1.4), we cannot know how much each of these abundances contribute to the total protein amount. Also, we do not know the distribution of shared peptides intensities across their associated proteins, thus complicating their integration in protein abundance.

Note that even to obtain quantification at peptide-level, an aggregation step is required, but it is not as challenging. In fact, during the MS analysis, a given peptide can be ionized in several ways (*e.g.*, with a charge +1, +2, etc.), resulting in separate identifications and quantifications for different ionized versions of the same peptide (the precursor ions). Fortunately, the precursor-to-peptide aggregation has much less impact on downstream analysis, than from peptide to protein. First, because most peptides have only one precursor, and only few have as much as three of them; and second, because there are no such thing as shared precursors. Hence, it has long been assumed that this aggregation can be done before peptide filtering or after imputation without dramatic changes on data.

Differential analysis

We now have an abundance table at protein-level, without missing values. The most common question answered by label-free LC-MS/MS experiments is that of differential analysis. It consists, for each protein, in testing whether the fold change of abundances between several conditions is significantly different [Wieczorek17], thus detecting proteins differentially expressed. We then obtain a large number of p-values, for which many of them may be significant by chance. Thus, similarly to the identification problem (see section 1.1.4), there is a need to define a p-value cutoff to control the number of proteins falsely considered as differentially abundant (in the statistical jargon, one has prevented an excessive probability of type I error). We will also discuss this issue in the next section. Yet, proteomics experts have interest in other types of downstream analyses, such as visualization of the data, or clustering of proteins according to their abundance pattern over the different conditions.

The match-between-run option

An optional, yet often used MS signal processing step is known as match-between-run (MBR) [Tyanova16] (also referred to as cross-assignment, depending on the software used [Bouyssié20]). The idea is to associate the quantification of a non-identified peptide in a sample A, when it has been identified and quantified in another sample B. Given its MS peak attributes such as m/z, charge state, and retention time in sample B, the objective is to locate a peak in the MS signal of sample A that most likely corresponds to the same peptide. Then, we can consider this peptide to be present in sample A and compute a XIC value from the peak retrieved. This option enables to reduce the number of missing values in the peptide table, based on real XIC abundances instead of statistically imputed values. However, while tackling the MV issue, MBR also increases significantly the number of falsely detected peak [Lim19]. Thus, a rigorous statistical procedure should be involved to limit their number. However, doing so is not straightforward as it cannot be included in the identification procedure (see section 1.1.4) because of the lack of matching MS/MS spectra, and further methodological developments are needed in this direction. Hence, although MBR can reduce the number of MVs, it can only do so to a certain extent, as an important number of MVs still needs to be tackled afterwards.

Methodological development at EDyP Lab

The entire label-free LC-MS/MS pipeline can be broadly decomposed in two main parts: the "wet-lab" processing, encompassing sample preparation, digestion, and the LC-MS/MS analysis; and the "dry-lab" processing, with all the computational steps (identification, quantification, and statistical analysis). For the "wet-lab" part, EDyP Lab uses commercial tools and products. However, most of

computational tools are in-house products, as the processing of high-throughput LC-MS/MS data is an active field of research. EDyP has for example developed Proline (a software tool proposing a variety of functionalities for precursor to protein aggregations and for quantification [Bouyssié20]) and Prostar (for the statistical analysis [Wieczorek19]).

The work of this thesis is part of a drive to develop robust and efficient statistical tools, which can be integrated to these in-house software tools. As both are freely accessible, we ambition their continuous improvement can benefit to the entire proteomic community.

1.1.5 Major statistical issues in proteomics

We address in this thesis work three major issues in the statistical processing of proteomic data, which are thereby described.

False Discovery Rate control

In statistics, the multiple testing problem arises when a large number of statistical inferences are made simultaneously from observed values. The more inferences are made, and the more frequent spurious inferences are. For example, let us suppose that researchers are testing several drugs individually to see whether they can cure an illness. If the drug efficiency is measured on a biological variable subject to stochastic variations on a too small number of patients, there is a chance that random fluctuations perfectly match the severity of the cohort patients, so that one ineffective drug is believed effective. Naturally, the probability of such event increases with the number of drugs: the more drugs tested, the higher the risk.

This issue occurs in two major steps of LC-MS/MS pipeline: the identification of peptides and the differential analysis. In the first case, the inferences correspond to the list of PSMs obtained from the search on the target database, with their associated score. In the second case, the inferences are the statistical tests and associated p-values comparing mean abundances for each protein. In both cases, we seek a threshold that would retain as many correct inferences as possible, while limiting the number of inferences retained by chance.

The false discovery rate (FDR) framework fits particularly well the issue here. It can be defined as the following. We establish a large number of hypotheses, referred to as null hypotheses (*e.g.*, "the protein abundances do not differ between two biological conditions"), and define for each the corresponding alternative hypothesis (*e.g.*, "the protein abundances differ between two biological conditions"). We then apply a procedure to select which null hypotheses are rejected in favor of the alternative one (a rejected null is called a *discovery*). This procedure often consists in computing a p-value or a score for each null, and the rejection is assessed when it falls beyond a specified threshold value t. Formally, the FDR at level t reads:

$$\mathrm{FDR}(t) := \mathbb{E}\left[\frac{\#\text{ of } \textbf{true} \text{ null rejected at level } t}{\#\text{ of null rejected at level } t}\right] = \mathbb{E}\left[\mathrm{FDP}(t)\right]$$

Where the expectation here is taken over the distribution of the data, and $\mathrm{FDP}(t)$ refers to the false discovery proportion (it can be viewed as 1-precision, in the statistical classification jargon), which is in practice unknown. Then, to control for the FDR, we need to have a procedure which gives the threshold t such that:

$$\mathrm{FDR}(t) \leq \alpha$$

where α is user defined risk value, such as for example, at 1% or 5%.

In differential analysis, the FDR control has long been relied on (*e.g.*, for RNA microarray type experiments [Efron02, Tusher01]), and has solid guarantees as the null hypothesis is usually well characterized. However, the null hypothesis for PSM at identification step is much harder to characterize: let it be “the PSM occurred by chance,” how do we define “by chance?” Even though the concept of FDR control is more recent for validating PSM than for testing the mean

equality between two groups, a lot of efforts have been put by the proteomic community into the development of appropriate methods. However, until recently, those methods lied on empirical rules without further theoretical guarantees. Recent theoretical advances in statistics could help justifying *a posteriori* these empirical rules, however, they also suggest modifying them. Therefore, a significant part of this thesis has been focused on unifying and landmarking these elements from distinct scientific communities.

General FDR control methods are described in section 1.2 and as well methods used in proteomics, both for differential analysis and identification of peptides.

The missing values issue

The important proportion of MVs, as well as the complexity of their nature, which cannot be determined (see section 1.1.4), drastically complicates their handling, and can strongly affect subsequent analysis. For example, if a peptide has many MVs in a given biological condition with respect to another, not accounting for the low abundance censoring can lead to biased estimation of means, and we may miss a reliable biomarker.

To the best of our knowledge, no concrete evidence exists regarding whether MV should be handled at precursor or peptide level. However, it is preferable to treat them before the aggregation of abundances to protein-level [Lazar16]. In fact, aggregating observed and non-observed peptide values with current pipelines often amounts to imputing by the neutral element of the aggregation rule employed (for example a zero in case of sum-based aggregation). This implicit imputation is not desirable as it risks distorting the protein signal, especially when few peptides are used in the aggregation.

Two main approaches can be distinguished to handle missing values in proteomics [Taylor22]. Firstly, one can simply keep the MVs as such during the data processing. This is possible when the data processing tools can handle MVs by relying on the observed values available only. For example, it is always possible to perform a t-test on the abundance values between two different conditions as long as at least two values are observed in the compared conditions, even though it may lack of power. Alternatively, one can impute these values, *i.e.*, replace MVs with plausible values regarding the structure of dataset, and then treat the dataset as a completely observed one, enabling any type of analysis. As doing so may lead to place too much confidence on "created data," a refinement of this approach is to perform multiple imputations [Chion22]; one runs a non-deterministic imputation algorithm several times on the same dataset, as to account for the uncertainty of the imputation. As a downside, it requires specific strategies to subsequently integrate them into a stand-alone result, which are discussed in subsection 1.3.2. In any case, the treatment of missing data in LC-MS/MS experiments is challenging as it can strongly affect downstream analysis, and no consensus has emerged yet on how they should be handled.

We define the missing values in proteomics from a statistical point of view, and present state of the art methods to handle them in section 1.3. We specifically focus on imputation, which is the method preferred by EDyP Lab, and present our main contributions on this topic.

Proteogenomics and transcriptomic integration

Proteogenomics refers to the joint study of proteomics and transcriptomics (or genomics) [Tariq21]. It falls within the scope of the so-called "multi-omics" approaches, which aim at combining different omics to extract more and better knowledge than with a single omic technology.

In gene expression studies, the combined use of proteomics and transcriptomics is of interest to leverage both the high coverage and precision of RNA-Seq technology, and the direct but often less complete measurement of protein abundances by LC-MS/MS. Proteogenomics typically involves enhancing proteomics analysis by utilizing the transcriptomic identifications or genome of the same sample to construct the target database, with the goal of improving protein inference [Miller22] (see section 1.1.4) or peptide identification [Fancello22]. While other studies have suggested integrating

transcriptomic expression data to improve peptide identification [Shanmugam14, Ma17], no study has yet focused on enhancing the quantification of proteome using transcriptomic analysis. Yet, variations in transcript levels between several conditions could give valuable information when dealing with missing peptide abundances. However, the integration of transcriptomics analyses to improve or extend proteomic quantitation poses considerable challenges, notably because of the complexity of both quantitative and qualitative dependencies between them two.

We discuss some works depicting the relationship between quantities of mRNAs and proteins, and we present state-of-the-art methods to extrapolate protein measurements from transcriptomics, as well as our contribution on proteogenomic imputation in section 1.4.

1.2 The control of FDR in proteomics

1.2.1 FDR control methods

The notion of FDR has been introduced in 1995 by Y. Benjamini and Y. Hochberg [Benjamini95] as a radically new solution to the classical problem of multiple testing problem. It has since then thrived into a large family of methods. We hereafter give an overview of most common ones.

Benjamini-Hochberg procedure

The Benjamini-Hochberg (BH) procedure [Benjamini95] was originally designed to control for the FDR using a list of p-values resulting from independent statistical tests. It consists in selecting a threshold beyond which null hypotheses are rejected, given a user probabilistic defined upper bound on the FDR. Further work [Benjamini01] demonstrated that the FDR remained controlled under positive regression dependency between test statistics, thus proving the procedure's efficiency for a great variety of applications. Also, the estimation of the overall proportion of null hypotheses among all of them, denoted π_0, can be integrated to the procedure, and enables a more precise FDR estimate [Storey03]. We next summarize the general patterns common to the most used BH-related procedures, as well as a bit of the underlying intuition.

Consider m hypothesis $H_1, \dots, H_m$ tested with their associated p-values $p_1, \dots, p_m$. Let us denote $p_{(1)}, \dots, p_{(m)}$ the sorted p-values by increasing order, and $H_{(i)}$ the null hypothesis associated to $p_{(i)}$. Let q^* be a user-defined risk threshold (often referred to as the *target FDR* by proteomists): it means one wants the FDR of our selection procedure to be lower than q^*, resulting in a so-called conservative control. Let us suppose we estimated an overall proportion of null π_0 among the m hypothesis (an intuition on how to do so is given later). Note that setting π_0 to 1 (as in the seminal BH article) will always result in a conservative procedure, although it is usually not optimal. Then, the BH procedure reads as:

$$\begin{aligned} &\text{Choose } k, \text{ the largest } i \text{ such that } \frac{p_{(i)}\pi_0 m}{i} \leq q^*, \\ &\text{Then reject all null hypothesis } H_{(i)} \text{ for } i \in \{1, \dots, k\}. \end{aligned} \tag{1.1}$$

Doing so ensures that $\mathrm{FDR}(p_{(k)}) \leq q^*$, but contrarily to a rather frequent misconception in the proteomic community, it does not ensure that the $\mathrm{FDP}(p_{(k)}) \leq q^*$. The quantity $\frac{p_{(i)}\pi_0 m}{i}$ can only be viewed as a "guess-timate" of $\mathrm{FDP}(p_{(i)})$ as illustrated on Figure 1.6. However, its minimum over the i's$\geq k$, which is $\min_{i:i\geq k} \frac{p_{(i)}\pi_0 m}{i}$ is an estimate of $\mathrm{FDR}(p_{(k)})$, also referred as a q-value [Storey02]. The FDR control is tighter when π_0 is well estimated and the p-values are well calibrated (*i.e.*, the p-values of true null hypotheses are effectively uniformly distributed on $[0,1]$), as illustrated on Figure 1.6 for a given p-value threshold $q^* = \alpha$. Broadly speaking, knowing the distribution of

p-values under null hypothesis, one can "subtract" this distribution to the observed one to have an estimate of the FDR. The uniformity of null p-values distribution also helps to correctly estimate the value of π_0: far away from rejection region (usually concentrated near 0), we expect to have only null p-values, therefore the height of the histogram on the right side should relate to π_0. Yet, p-values are in practice almost never as well calibrated as on Figure 1.6, so that more robust methods have been proposed to estimate π_0 (e.g. [Storey03, Storey04, Pounds06]), and some of them are routinely used in proteomics [Giai Gianetto16].

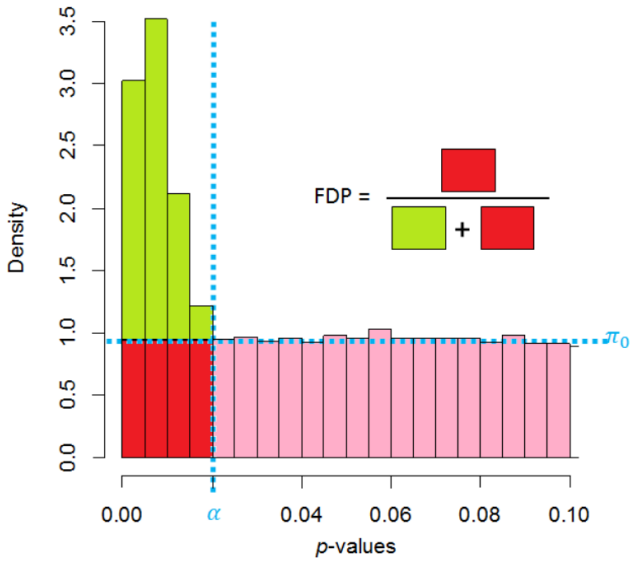

Figure 1.6: Illustration from [Burger18] of the estimation of FDP in the BH procedure. The red area corresponds to the density of p-values from true null hypothesis, while the green area corresponds to the one from true alternatives.

Empirical Bayes approaches

According to the definition above, the BH procedure is limited to situations where p-values are available. Empirical Bayes approaches [Efron02, Efron01] have enabled to extend it to other types of scores, not even necessarily related to a statistical test. It consists in estimating the distribution of scores under the null hypothesis, before its subsequent "subtraction" from the complete distribution. This estimation can be performed using parametric assumption on the data distribution, or via an alteration of observed data signal, such that the distribution of resulting scores distribute like under the null, for example by randomly permuting samples [Efron01].

Similarly to the above example, let us take a list of ordered scores $y_{(1)}, \dots, y_{(m)}$. In this example we consider the lower the score is, the less the null hypothesis is plausible. Thus, we also aim at finding a threshold below which we reject null hypotheses, given a target FDR q^*. Let us denote $\bar{F}$ the empirical cumulative distribution function (c.d.f.) of the scores $y_{(i)}$ and F_0 the cumulative distribution function of the null scores. In this setting, we can control for the FDR at level q^* by tuning the rejection threshold to the largest $y_{(i)}$ such that:

$$\frac{\pi_0 F_0(y_{(i)})}{\bar{F}(y_{(i)})} \leq q^* \tag{1.2}$$

Note that, when the distribution of null scores is exactly known, this method is akin to the BH approach, as the term $F_0(y_{(i)})$ can be viewed as the p-value resulting from the test statistics $y_{(i)}$, knowing then that $\bar{F}(y_{(i)}) = i/m$. An important advantage of this approach is that arbitrary scores

can be used, as long as it is possible to approximate the associated null distribution. This can be done in a pseudo-generative manner: for example, let us consider a table with thousands of gene expressions. For each gene, if one randomly shuffles its values across all samples, the distribution of statistics assessing difference of mean between conditions should be the one from the null hypothesis (*i.e.*, there is no mean difference). Then, one approximates F_0 by the c.d.f. of the statistics obtained [Efron01].

Knockoff filter

More recently, a new framework has been proposed for FDR control, called *knockoff filters* [Barber15, Candès18]. It was originally designed to control for the FDR in a variable selection task on tabular data, as to select a subset of features from a set $X = (X_1, \dots, X_m)$ that best explains an outcome variable Y (*e.g.*, a severity state of a disease.) The null hypothesis for a given variable X_i can then be stated as: Y and X_i are independent conditionally to the other variables $X_{-i} = X \setminus \{X_i\}$. To control for the FDR resulting from testing each variable, this method aims at generating so-called knockoff variables, *i.e.*, a set of variables $\tilde{X}$ of the same size as X, which respects two properties:

1. *Exchangeability:*
$$\forall S \subset \{1, \dots, m\}, (X, \tilde{X}) = (X, \tilde{X})_{\text{swap}(S)},$$
where the equal sign stands for equality in distribution, and $(X, \tilde{X})_{\text{swap}(S)}$ is obtained from $(X, \tilde{X})$ by swapping the entries X_i and $\tilde{X}_i$ for each $i \in S$.

2. *Conditional independence:* $\tilde{X}$ and Y are independent conditionally to X.

While the second property ensures that variables generated are truly under the null hypothesis, the first one is crucial for the following step. One then computes a contrast score W_i for each $i \in \{1, \dots, m\}$. A high positive (resp. low negative) contrast score W_i indicates that the variable X_i is more (resp. less) explanative of the outcome variable Y than its original knockoff counterpart $\tilde{X}_i$. A necessary condition on these scores for the procedure to work is that that the contrast scores of variables under the null hypothesis distributes symmetrically around 0. The exchangeability of knockoff variables aims at fulfilling this symmetry condition. Therefore, the contrast score must be carefully defined accordingly.

In any case, we should keep in mind that simply copying the original data distribution to generate knockoffs will result in poor statistical power (*i.e.*, high type II error), as the requirement 2 of knockoff definition would be poorly fulfilled. The more the knockoff variable is independent from its original variable, the more power we can expect, and this is what the proposed methods attempts to do.

Finally, to control for the FDR at a given target value q^*, one rejects the null hypothesis for variables with a contrast score greater than a threshold t, defined as:

$$t = \min \left\{ \tau > 0 : \frac{1 + \#\{i : W_i \leq -\tau\}}{\#\{i : W_i \geq \tau\}} \leq q^* \right\} \tag{1.3}$$

After the seminal work of Barber and Candès, multiple methods have thrived to generate knockoffs from various data types [Candès18, Romano19, Kurz22, Sesia19] as to improve this new approach to FDR control. Indeed, the resulting type of estimators strongly differs from the BH and empirical Bayes related methods, as here the selection is made according to contrast scores with different properties than score simply reflecting the importance of a variable. In fact, the term $\frac{1+\#\{i:W_i \leq -\tau\}}{\#\{i:W_i \geq \tau\}}$ can be viewed as a conservative estimate of $\text{FDP}(\tau)$ [Candès18], as the term $\#\{i : W_i \leq -\tau\}$ is equal in theory to the number of null variables with contrast score above τ, and the $+1$ term ensures we are on average above the FDP.

1.2.2 Applications of FDR control methods to proteomics

In LC-MS/MS proteomic experiments, FDR control is required both at peptide identification step and differential analysis. However, methods dedicated to control false peptide identifications have been developed and have spread based on empirical arguments only, and efforts to theoretically support them have long remained scarce.

FDR control for identification of precursor peptides

We recall first the multiple testing issues encountered at the identification step. Each MS/MS spectra obtained is used to query a reference *target* database of theoretical spectra that can be found in the protein mixture analyzed. The best match of the MS/MS spectra onto the target database is referred as a peptide-spectrum match (PSM), which is a triplet {MS/MS spectra, target sequence matched, matching score}. Regarding the large number of PSMs obtained, as well as the varying quality of the MS/MS spectra, we expect a non-negligible number of PSMs to result from a random match, and thus to be devoid of biological validity. As such random matches are expected to be less frequent when the match score is high than when it is low, it makes sense to define a cutoff score to retain the sufficiently confident PSMs only. In a multiple test correction parlance, the null hypothesis for a given PSM is that the assignment between the MS/MS and the target one is random (and thus incorrect). We then aim at controlling for the FDR by selecting a rejection region on the range of PSM scores.

To do so, the most common methods require a so-called *decoy* database. It is alike the target database except that it contains amino acid sequences that should not be found in the sample (and which do not appear in the target database). These sequences are often generated to mimic the target database, for example by reversing the target sequences [Moore02], shuffling the target sequences [Klammer06], generating the decoy sequences at random using a Markov model with parameters derived from target sequences [Colinge03], or by other methods. Yet, no clear consensus on how they should be generated has been established, neither on the choice of the decoy database size (which for simplicity is often equal to that of the target database) [Jeong12]. This decoy database can be used in two different manners (see Figure 1.7), which forms the two main approaches to FDR control that have been used, confronted, and compared in the proteomics literature for the last 15 years.

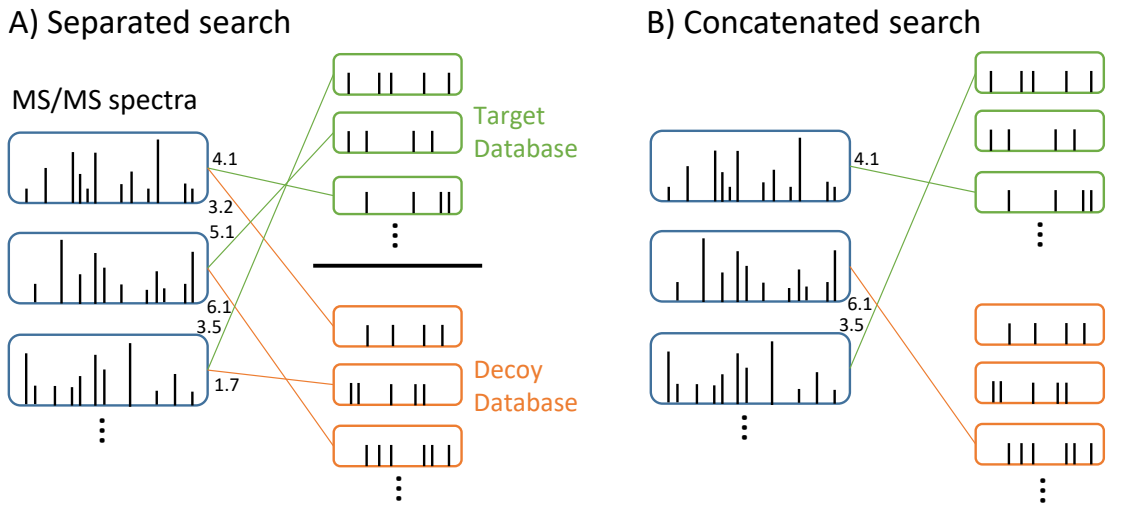

Figure 1.7: Illustration of separated (A) vs concatenated (B) database searches when using decoy database to control FDR at identification step. Each line represents either a target PSM (green) or a decoy PSM (orange) in the search setting given.

The first one consists in applying the search procedure twice: one on the target database, and another one on the decoy database [Moore02, Käll08]. This is often referred to as a *separated* search. The second one consists in concatenating the target and decoy databases, and operating the search on the resulting concatenated database [Peng03, Elias07a]. In this setting, for each MS/MS spectrum, the target and the decoy databases "compete" to propose the best match, which is why *concatenated search* is also often referred to as *target-decoy competition* (TDC).

Both methods share features: First, they rely on an important assumption referred to as the Equal Chance Assumption (ECA). It states that incorrect matches on the target database and decoy matches are equally probable. Although intuitively grounded, the ECA can hardly be assessed, and strongly depends on how the decoys are generated.

Second, both methods have inspired creative researchers with a biology or biochemistry background, who have proposed many FDR guess-timates based in the counts of target and decoy matches passing a given threshold. Among them, few should nonetheless be retained, as they were more motivated by improving the statistical rigor than by artificially inflating the rejection region. Notably, the ratio D/T, where T (respectively D) is the number of target (respectively decoy) PSMs passing the significance threshold, has been first proposed in separated search setting [Käll08], whereas $(D+1)/T$ has only recently been proposed for TDC [Levitsky17].

Controlling for incorrect peptide identifications is still an active field of research, and new methods are continuously emerging. Except from [Couté20], which authors proposed to bypass decoy generation by directly relying on the properties of the search engines scores (as to recover p-values for subsequent application of BH procedure), most elaborate on more refined uses of the target and decoy databases. Among them, one is worth mentioning, as it tightly relates to the work present in chapter 2: Madej and Lam have proposed in [Madej22] a general null PSM score distribution (based on the empirical Bayes FDR framework) estimated from plenty of publicly available datasets. The idea is that this null distribution, referred as Common Decoy Distribution (CDD) is not dataset-specific, but only depends on search engine and on some search parameters.

Differential analysis

Differential analysis is a task common to many omic approaches, and it has long relied on FDR theory. The first FDR control methods, notably BH and empirical Bayes, have long found practical applications in Genome Wide Association Studies (GWAS) or RNA microarray data [Sun06, Efron02]. When proteomics emerged, it naturally borrowed this know-how, which explains why they are now used in most proteomic experiments. Yet, it should be noted that other recent methods enable to incorporate prior knowledge or informative covariates to prioritize, weight and group the hypothesis tested (detailed review in [Korthauer19].) This can be useful in proteomics as for example, the number of identified peptides or the number missing values can affect our confidence on a detected biomarker [Burger18]. It can lead to more accurate and "personalized" q-values for each protein.

1.2.3 Beyond FDR: variable selection for patient diagnosis

FDR control and more generally multiple testing correction methods are common approaches to propose new biomarker candidates, as they make it possible to quantify and control the risk that a candidate appears to be a false positive. These are univariate approaches that do not cope with dependencies between variables. However, combining several variables can yield more powerful biomarkers for complex biological status (*e.g.*, to diagnose the state of a patient), and to do so, classical FDR procedures are limited. For example, FDR control can lead to discover several biomarker candidates, but, if these are well correlated, their combination in a multivariate model (often referred as a "panel" in clinical context) will not necessarily increase prediction performances with respect to a univariate model. However, there are popular variable selection methods such as Lasso [Tibshirani96] or Elastic Net [Zou05] that enable to select the set of variables that explains

the best, altogether, an outcome variable Y. These methods require hyperparameter tuning, which can be unstable when too few data are available, and unlike the FDR approach, do not address any quality control requirements.

To compare panels, generalized linear models (GLMs, which are flexible generalization of ordinary linear regression) are insightful, as they enable comparison between several nested models [McCullagh19]. A GLM A is said to be nested in another GLM B if all its variables are included in those of B. Hence, a nested comparison test between A and B fitted on the same dataset, consists in testing whether additional variables included in B bring significant improvement to the likelihood or not. These tests are particularly convenient when a small number of variables is considered, and when prior knowledge about some of them being efficient is available, as then, the combinatory remains low. Therefore, FDR approach can be used to select a small subset of biomarkers from a large list of covariates, and nested hypothesis tests would then help to decide which combination of covariate has the highest predictive performances. However, doing so requires being careful about data leakage between the various steps, as to avoid overfitting [Desaire22].

1.2.4 Contributions

We sum up here our main contributions related to FDR control in proteomics, which are thereby detailed in chapter 2:

1. **Univariate contrast scoring for knockoff procedure:** In section 2.2, we apply knockoff procedure on proteomic quantitative data for differential analysis. To cope with high dimensionality of the data, we adapt the knockoff procedure by proposing a univariate p-value based contrast scoring method.

2. **Novel theoretical considerations on TDC enlighten by knockoff filter:** Theoretical links between TDC and knockoff procedure have been identified by the proteomic community. Whereas they were often used to justify the TDC approach, we present in section 2.3 some important discrepancies. Regarding those, some limitations of the TDC approach naturally appear, and we propose perspectives on how they could be overcome.

3. **Development of a diagnosis panel from biomarker selected with FDR control:** We propose an original way to combine previously discovered biomarkers with new ones discovered in a study performed at EDyP Lab, to establish a multivariate diagnosis score. This highlights a practical application following up biomarker discovery under FDR control.

1.3 Missing Values in Proteomics

1.3.1 Statistical considerations on proteomics MVs

Rubin proposed more than 40 years ago a statistical classification of the mechanism underlying missing values [Little19]. We define them here formally. For simplicity and without loss of generality, let us have a complete random vector X containing the information of interest. We denote M the random vector associated to the missing response for each entry of X (containing a 0 for an observed entry and 1 for a missing one), and subscripts “obs" and “mis" respectively the set of indices of observed and missing entries in X. We refer as *missingness mechanism* the conditional probability distribution $P(M|X,\Gamma)$ where Γ denotes unknown *missingness parameters*. Then the missingness mechanism falls in one of these three cases:

- **Missing Completely at Random (MCAR)**: the probability for an entry to be missing does not depend on any other entry:

$$P(M|X,\Gamma) = P(M|\Gamma) \tag{1.4}$$

- **Missing at Random (MAR)**: the probability for an entry to be missing may only depend on observed entries:

$$P(M|X,\Gamma) = P(M|X_{\text{obs}},\Gamma) \tag{1.5}$$

- **Missing Not at Random (MNAR)**: any other case, *i.e.*, when the probability for an entry to be missing depends on missing entries.

MVs on which MNAR mechanism applies (that we refer to as MNAR values) are the most challenging as we do not assume any simplification over the dependence between M and X. It thus means we need to account for the missingness response of entries to infer conclusions from the dataset. We illustrate this issue with a situation where one wants to estimate the parameters Θ describing the complete data X by a maximum likelihood estimation (MLE). In the presence of missing values, a natural way to estimate Θ is to maximize the joint likelihood of the entire data available, *i.e.*, of the joint observations of (X_{obs}, M). Indeed, more than X_{obs} is available, as the knowledge of M and of its possible dependency to Θ is informative too. Let us suppose then that Θ and Γ are distinct, *i.e.*, their joint parameter space is the cross product of each of their parameter space. In an MAR or MCAR setting, it is straightforward to show that [Josse18]:

$$P(X_{\text{obs}}, M|\Theta,\Gamma) = P(M|X_{\text{obs}},\Gamma)P(X_{\text{obs}}|\Theta) \tag{1.6}$$

This result shows that, in such setting, we can find an MLE of Θ by simply maximizing the likelihood of the observed values X_{obs}; as $P(M|X_{\text{obs}},\Gamma)$ does not depend on Θ. Oppositely, in the MNAR setting, the latter term would depend on Θ, so that it is necessary to account for the missingness mechanism to perform MLE.

In LC-MS/MS proteomics, we usually assume there are two missingness mechanisms at stake [Lazar16, Webb-Robertson15]. Missing values that occur randomly because of some unexpected issues during the complex LC-MS/MS pipeline, such as mis-cleavage, mis-identification, or other (see section 1.1.4 for more details), are considered MCAR. On the other hand, missing values related to the dynamic range of view of the instrument, low abundance, or even the complete absence of the peptide in the sample, are considered MNAR.

No convincing MAR mechanism has been identified yet. In a dataset of peptide abundances, such mechanism would occur if the missingness of a peptide would only depend on the missingness of other observed peptides. A recent paper [Gardner21] claimed that MVs due to peptide misidentification (leading to an observed but wrongly assigned value in another peptide) can be considered MAR. This statement seems to lack support as the missingness response of the missing peptide mostly depends here on the missingness response of the peptide carrying the wrongly assigned value, and not the value itself.

To handle MNAR values, Little and Rubin [Little19] defined two approaches, corresponding to two different factorizations of the joint distribution of M and X:

- **The selection model:**

$$P(X,M|\Theta,\Gamma) = P(X|\Theta)P(M|X,\Gamma)\,, \tag{1.7}$$

  where we retrieve the distribution of the entire distribution of X in the first factor, and the missingness mechanism in the second.

- **The pattern-mixture model:**

$$P(X,M|\Psi,\Omega) = P(X|M,\Psi)P(M|\Omega)\,, \tag{1.8}$$

where the first factor characterizes the distribution of X given the pattern of missingness, parameterized by Ψ, and the second refers to the marginal distribution of missingness response defined by parameter Ω (again we assume Ω and Ψ are distinct).

The choice between these two models depends on the type of data and on the application encountered [Little19]. The selection model seems natural when we are interested by dependencies between covariates of X over the whole dataset. On the other hand, the pattern-mixture model often provides more interpretable parameters for subject-matter experts, and is thus particularly used in sensitivity analysis, which aims at evaluating the uncertainty of the output of a model linked to few parameters. However, strong restrictions on the parameter space are always necessary in the pattern-mixture approach [Little19].

Most of the work modelling the distribution of LC-MS/MS proteomic missing data rely more or less explicitly upon the selection model [Karpievitch09, Luo09, Ryu14, Chen14, O'brien18, Li23], as the estimation of the missingness mechanism, related to the instrument quantitation limits, seems to be a natural approach to understand the underlying overall distribution of the data. Among them, some works [Chen14, O'brien18, Li23] do not model the presence of MCAR values and instead consider only a global random left censoring mechanism, for example with a Probit (the normal cumulative distribution function) or a Logit (the logistic function) mechanism. This choice is simpler and may not be far from the experimental reality, as, in any case, we do not have any direct way to know whether a value for a given peptide is MCAR or MNAR. Also, a selection model that integrates a Probit or Logit missingness mechanism is guaranteed to be identifiable under some assumptions [Miao16]. Identifiability means here that two different parameters cannot lead to the same joint distribution (X_{obs}, M), and it is thus a desirable mathematical property for parameter estimation.

1.3.2 Handling MVs in proteomics

In LC-MS/MS proteomics, *complete-case analysis* [Little19], *i.e.*, filtering out all peptides containing MVs and conducting data analysis subsequently, is not considered. As the proportion of peptides having at least one MV in as dataset often reaches more than 60% [Liu21], complete-case analysis would discard way too much information. Also, because of the expected large amount of MNARs, any subsequent analysis of the data would be highly biased (cf. subsection 1.3.1). Yet, in practice, peptides with too few observed values are filtered out, as we consider they may not be reliable in the subsequent analysis.

Another approach, referred to as *available-case analysis* [Little19], consists in leaving missing values as such, and conducting subsequent analysis, nonetheless. The corresponding methods will be referred to as *"imputation-free,"* and naturally depend on the objective(s) of the analysis. For example, one can compute a variance and a mean when two values are observed in each biological condition, and then use them to perform a t-test. One of the most popular tools for differential analysis in gene expression tables, *limma* [Ritchie15], can natively handle MVs, as long as at least one value is observed in at least two conditions. However, it does not cope with the presence of MNAR values and would thus return biased conclusions on LC-MS/MS proteomics data. On the other hand, numerous tools have been developed to handle MNARs in proteomic differential analysis. Among them, some [Ryu14, O'brien18] rely on similar Probit missingness mechanisms to determine the mean peptide abundances in each condition, with Bayesian or MLE based estimators. The tools MSqRob [Goeminne20] relies on an orthogonal approach, as it aggregates p-values from differential expression and differential detection (testing whether a peptide is more detected in a condition than in another) to assess whether the peptide is relevant or not. Finally, a recent and yet unpublished method [Chion23], uses a Bayesian framework that leverages the intra-condition correlations between the peptides resulting from same proteins to assess the posterior distribution of the mean and variance in each condition. These posterior distributions can then naturally be used for differential analysis. This latter method is, one of the first attempts to leverage dependencies

between peptides from same proteins to improve quantitative analysis, which is a path we have also focused on (see chapter 3).

The clear advantage of imputation-free methods with respect to imputation methods, whether there are used to infer parameters or for differential analysis, is that they avoid any risk of data distortions. For example, imputing MVs that are known to be left-censored by a too low value may artificially increase peptides variances in some conditions, resulting in a loss of power in differential analysis. Following an inverse logic, imputing missing values by the observed mean of their respective condition is probably the worst thing to do, as the intra-condition variance would shrink, and hence most of imputed peptides would appear as significant. This latter example also raises the question on whether the imputation should be done by accounting for the experimental conditions, which is an issue we discuss in section 2.5. Yet, a main advantage of imputation is that it enables any type of downstream analysis. In fact, if properly done, there is no need to create, for each type of downstream analysis, a model or pipeline dealing specifically with random left-censored MNARs. We distinguish here three different frameworks regarding missing value imputation in LC-MS/MS proteomics.

First, **single imputation** consists in imputing a dataset once (with one method), and then in conducting the analysis as if the dataset had been fully observed. The imputation methods used to do so can often be classified between MCAR/MAR-devoted methods and left-censored oriented methods (hereafter referred to as MNAR-devoted methods for simplicity). We provide a selective review as well as references towards more exhaustive reviews in the following section.

Second, we call **meta-imputation** any method that consists in aggregating results from several single imputation methods. The underlying motivation is often that combining MCAR and MNAR values should limit imputation bias, either by interpolating the results [Ma20], by iterative sampling with different imputation methods [Wei18, Wang22] or by selecting the appropriate method for each MV [Giai Gianetto20, Gardner21]. Among the latter, let us mention the *imp4p* package [Giai Gianetto20]. It proposes several meta-imputation strategies, all based on a preliminary diagnosis: for each MV, it estimates the probability of being either MCAR or MNAR. Then, this probability is used to refine the combination of several imputations accordingly. Note that *imp4p* directly falls in the scope of the pattern-mixture model, where we aim at characterizing the marginal distribution of each pattern (MCAR, MNAR and observed values).

Last, we refer to as **multiple imputation** methods, any method that imputes several times a dataset and store the resulting multiple imputations for subsequent data analysis. The multiple imputations can either be done by different algorithms (for example MCAR- and MNAR-devoted ones) and/or with a stochastic algorithm that is given different seeds as input. Only few works follow this direction in proteomics. A well supported approach [Chion22] relying on Rubin's rule [Little19] has recently been proposed with an interesting feature: it proposes to include the variability of the multiple imputed values in the estimation of parameters (such as intra-condition mean and variance), as to directly cope with the variability due to the imputation-related uncertainty in the subsequent data analysis.

Finally, we have noticed one article which lies between imputation-based approaches and imputation-free ones, and which was an important source of inspiration for this doctoral work: PEMM (a penalized EM algorithm incorporating missing data mechanism [Chen14]) aims at estimating the overall mean and covariance matrix of peptides by considering an original missingness mechanism. As the missing values are the latent variables of the model, they are generated at each E-step, so that, according to the authors, it yields a natural imputation algorithm. Unfortunately, the authors did not benchmark the imputation capabilities and focused on the parameter estimation under incomplete data instead. The few attempts to use PEMM for imputation explicitly that can be found in the literature ([Hediyeh-zadeh23, Kong23], confirmed by our experiments) let us think it is indeed not adapted for this task, for a variety of reasons, among which, convergence issues, lack of scalability, and the necessity to manually tune the missingness parameter. Other limitations are discussed in detail in chapter 3.

1.3.3 State of the art on single imputation

Among the several options available regarding imputation, our work focuses on single imputation for several reasons. From a practical point of view, single imputation is the most straightforward and simple method to deal with MVs. Likewise, it is also the standard approach in most labs worldwide, including EDyP, thanks to its easy-of-use and versatility with respect to downstream data analysis. Therefore, this focus makes our work easy to integrate in most homemade statistical pipeline. Also, from a methodological point of view, the quality of the results obtained by meta or multiple imputation approaches directly depends on the quality of the single imputation method they stem from. Thus, the improvement of single imputation methods, adapted to LC-MS/MS proteomics data is an inescapable long term issue. We give here an overview of imputation methods that are commonly used or benchmarked in proteomics experiments, with first MCAR/MAR- and then MNAR-devoted methods.

MCAR/MAR-devoted methods include first methods based on dimensionality reduction, using a variety of state-of-the-art algorithms: Singular Value Decomposition (SVD) [Troyanskaya01], Probabilistic Principal Component Analysis (PPCA) [Ilin10] or Bayesian Principal Component Analysis (BPCA) [Oba03]. They all consist in finding a low dimension latent decomposition of the dataset, considering the dimensions removed are not relevant (as essentially containing noise, which in some case, can be explicitly modeled, as with PPCA or BPCA). Also, methods based on similarity between peptides are often considered, such as K-nearest neighbors (KNN) [Troyanskaya01], Sequential-KNN (SeqKNN) [Kim04], or Local Least Squares (LLS) [Kim05]. They first consist in finding the K closest peptides of the peptide to impute, according to a certain metric (Pearson's correlation, Euclidian distance *etc.*), and then either aggregating results by a weighted average (KNN) or by fitting linear models on each neighbor (LLS). At first glance, these methods do not seem suited for high dimensional setting, as the chances of finding apparently close peptides based on few observed values, and without any true dependencies, are high. This issue is yet partially addressed by SeqKNN [Kim04] which performs KNN iteratively from the peptides with the fewest MVs, to those with the largest amount. Finally, ImpSeq [Verboven07], a method that iteratively estimates the sample-wise covariance matrix and imputes missing values accordingly, has recently been introduced in comparisons benchmarks as it showed very good results. This approach is quite singular as contrarily to most imputation methods, it does not assume that the samples are independently distributed. However, this approach makes sense as most proteomics experiments involves several replicated samples per phenotype.

Most of these MAR/MCAR methods have been developed primarily for other gene expression studies (RNA-seq or microarray data), and usually report good performances on them. Although they capture global or local dependencies in the dataset, they do not account for the presence of left-censored MNAR values, which makes them ill-suited for LC-MS/MS data.

MNAR methods, on the opposite, are rather simple and empirical methods, which are most of the time univariate. For example, LOD (Limit of Detection) [Lazar15], simply imputes by the lowest value (or a low quantile of the distribution) in the sample or in the peptide; MinProb a slightly refined version, samples from a normal distribution around this LOD. QRILC (Quantile Regression Imputation of Left-Censored data) [Lazar15] first estimates the parameters of a Gaussian distribution by quantile regression and then imputes by sampling from this normal distribution truncated at LOD. A refined version of QRILC, named IGCDA (Imputation under a Gaussian Complete Data Assumption) [Giai Gianetto20] moderates the variance of impute values of QRILC. MsStats [Kohler23] relies on an accelerated failure time model [Taylor13], considering all missing values as censored by a fix detection threshold, in a similar manner as QRILC. However, few MNAR methods are not univariate.

First, trKNN (truncated KNN) [Shah17] is similar to the KNN approach, but instead of using the observed mean and variance estimates to compute the peptide correlations, it estimates these parameters by fitting a truncated normal distribution of each peptide with MVs. The choice of the

truncation point is however empirical and fixed to the minimum value observed in the dataset. A latter and recent method called msImpute [Hediyeh-zadeh23] has shown promising results. Briefly, it applies a low-rank SVD factorization on each condition independently, and then interpolates these results with a MinProb type algorithm, where interpolation weights depend on the intra-condition MV ratio. Thus, msImpute should adapt to varying missingness mechanism in the data. Here again, we retrieve implicitly the pattern-mixture model, as imputed values are a weighted average of different missingness patterns, corresponding to the number of MVs in a given condition. However, we have witnessed an issue with this interpolation approach: it does not always increase the performances of the algorithm (see chapter 3), so that adaptability to various missingness mechanisms is in fact limited.

There exists numerous reviews comparing the aforementioned imputation tools on LC-MS/MS proteomics data [Karpievitch12, Webb-Robertson15, Lazar16, Jin21, Liu21, Shen22]. However, no general consensus has emerged yet on which one to use, and conclusions regarding the best methods are often contradictory. The reasons behind this may be two-fold. First, apart from msImpute, no single imputation method explicitly copes with different missingness mechanisms, although they can vary from a dataset to another. Second, the validation procedure used for benchmarking often varies from one review to another, as there is no global consensus on the gold standard validation method [Harris23].

The validation of proteomics data imputation algorithms generally consists in two main approaches. The first one is mask-and-impute validation. It is based on introducing "pseudo-MVs" in a dataset with a chosen missingness mechanism, imputing the whole dataset, and computing the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE) or the correlation coefficient between the ground-truth and imputed values for each imputation method (see Figure 1.8 as well as subsection 3.6.5 for the RMSE/MAE definitions). Most of the related works have proposed mechanisms that controls the proportion of MNARs and MCARs among the masked values [Lazar16, Jin21, Wang22]. Thus, the sensitivity with respect to this proportion can be evaluated. Yet, the issue with these experiments is that we can draw conclusions only regarding the type of MNAR mechanism (Probit, Logit, non-parametric *etc.*) chosen for validation, as the true one remains unknown.

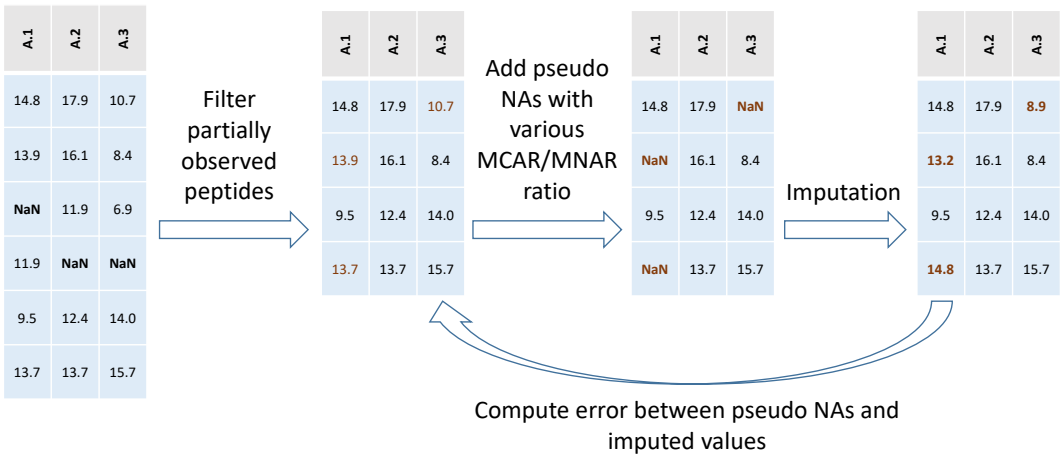

Figure 1.8: Description of typical mask-and-impute validation procedure.

The second main validation approach is oriented towards differential analysis. It consists in using benchmark datasets resulting from artefactual proteomic experiments. They are built as following: one spikes few human proteins (generally, one of the so-called Universal Protein Standards, or UPS, proposed by Sigma-Aldrich) at different (*i.e.*, varying across samples) known concentrations into a constant (*i.e.*, stable across samples) yeast or *E. coli* proteome, referred to

as the background; as to mimic a situation where only a few known (*i.e.*, labelled with a ground truth) proteins are differentially abundant. Technical replicates are produced for subsequent LC-MS/MS analysis and summarized in an abundance table. It has become customary to use those benchmark datasets to compare imputation strategies, by applying the following procedure: (1) impute the abundance table with different methods, (2) test each peptide for differential abundance, and (3) assess for each method whether low p-values correspond to UPS peptides with a ROC (Receiving Operating Characteristic) or (Precision-Recall) curve (see Figure 1.9). In addition, some authors [Liu21, Hediyeh-zadeh23, Harris23] assess whether the UPS fold-change (*i.e.*, the difference between mean abundances) computed after imputation agrees with the experimental procedure. A main advantage here is that validation relies on a known ground-truth and well controlled experimental conditions. However, this differential analysis validation has its limits, as these benchmark datasets are oversimplified and often unrealistic. Moreover, it does not ensure that well-performing methods in this context would remain so for other types of downstream analysis.

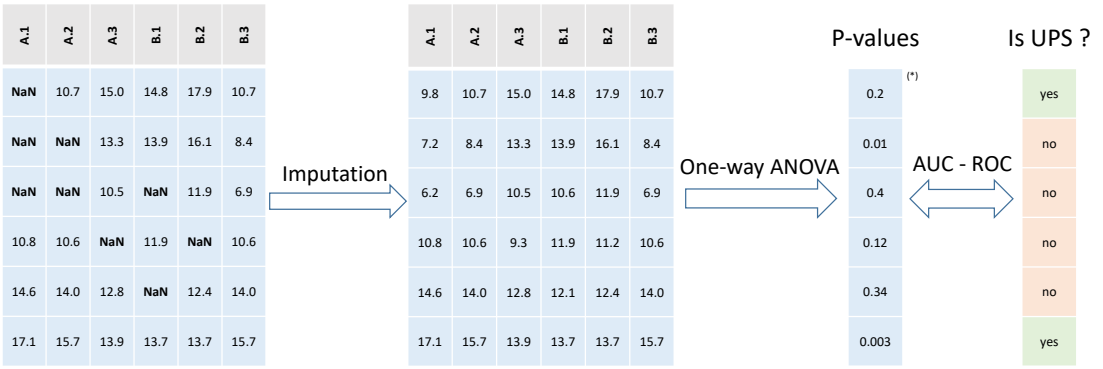

Figure 1.9: Description of differential-analysis oriented validation procedure. UPS here stands for peptides that are ground-truth differentially abundant.

To sum up, no global consensus exists on the standard validation procedures in proteomics, mainly because of the presence of MNAR values. A recent work [Harris23] suggested validations should almost entirely rely on benchmark datasets and account for other types of features which are interesting for downstream analysis (detection limit, peptides that are usable for protein quantification, runtime, etc.). However, owing to the simplistic natures of benchmark datasets, and to the importance of the imputation error, we disagree with this view.

1.3.4 Contributions

Our contributions to the missing value problem in discovery proteomics, thereby detailed in chapter 3, can be summed up as following:

1. **Development of a new imputation algorithm:** We developed a novel and original imputation algorithm named Pirat (Peptide or Precursor level imputation under random truncation). Pirat is easy to use and only requires few samples and at least one observed value per peptide (although keeping such peptide is often questionable). Finally, it has been implemented in an R package available on GitHub (`https://github.com/prostarproteomics/Pirat`).

2. **Pirat copes with biochemical dependencies:** Pirat drastically reduces the dimensionality of the imputation problem, by accounting for the dependencies between peptides or precursors derived from the same proteins, which has never been proposed in proteomic imputation before.

3. **Modelling the missingness for imputation:** Pirat's underlying model is based on Rubin's selection model. A great advantage of Pirat is that the missingness mechanism is automatically inferred from the dataset, enabling a better generalization of the method. Also, it does not require hyperparameter tuning as the latter ones are automatically set without additional computational costs.

4. **Pirat obtained the best performances** on benchmark datasets with endowed ground-truth about differentially abundant peptides. It also achieved best performances on mask-and-impute experiments with significant (and realistic) MNAR proportions.

1.4 Multi-omic integration for proteomics

1.4.1 Types of multi-omic applications in literature

The term "multi-omic integration" in quantitative gene expression data stands for a wide range of works, with different purposes [Rohart17, Argelaguet18, Song20, Tariq21]. Without pretending to be exhaustive, we mention here some of them:

- Discovery of biological pathways, or enrichment of previously discovered ones,
- Gene-set enrichment analysis, which aims at identifying classes of genes or proteins that are over-represented in some phenotypes,
- Classification of samples or individuals,
- Inference of an omic modality in an individual or sample from other omic modalities, which we thereby refer to as "sample extrapolation,"
- Inference of a whole unseen feature in a given omic modality from other omics modalities, which we thereby refer to as "feature extrapolation,"
- Missing value imputation,
- In general, compensation of weaknesses and limitations of different omic modalities.

The proteogenomics approach defined in section 1.1.5 globally falls in the range of the three latter points, as it aims at increasing the reliability and exhaustiveness of proteomic analyses by integrating transcriptomic and genomic information. We will thus focus on missing values imputation and sample/feature extrapolation for proteomics data from transcriptomics data in the rest of this section (description in Figure 1.10). Note that sample or feature extrapolation can directly be used as an imputation method, although it is intrinsically suboptimal, as it does not use the information available in the sample/feature that one seeks to impute. Moreover, we also restrict to cases where transcriptomics expression table of same or related samples are accessible, as well as optionally, the relation graph between genes, transcripts, and proteins.

1.4.2 Quantitative relationship between a transcript and a protein

To infer protein levels from the knowledge of transcript levels (where the term *level* refers here to a general quantity, either absolute or relative), we must assume there is a dependence between the two. This is particularly difficult to model at the scale of a single cell, as discrepancies may occur between these two levels across time (*e.g.*, the difference between half-lives of proteins and mRNAs, delay of translation process, *etc.*, see [Liu16]). However, a review based upon numerous works related to

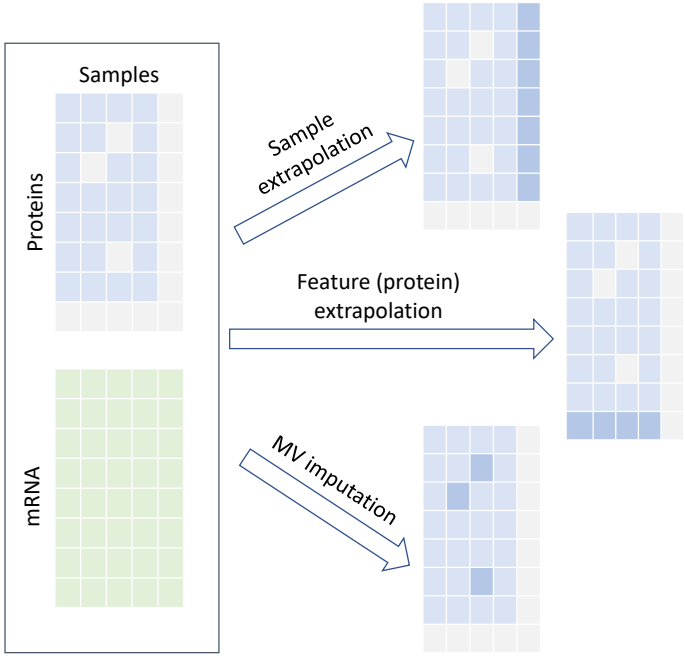

Figure 1.10: Possible quantitative integrations of transcriptomics data to increase coverage of proteomics data.

mRNA-protein quantitative analysis concludes that the number of protein copies in a bulk sample in steady-state conditions (*i.e.*, the overall number of protein/mRNA copies remains relatively stable across time) should be primarily determined by the number of transcripts copies [Liu16]. More particularly, it concludes that the difference between the number of protein copies across different genes is relatively well explained by the difference between the number of mRNA copies of the different genes. This means we can expect high gene-wise correlations between proteins and transcripts abundances (see Figure 1.11). Other authors clearly demonstrate this over various human tissues [Edfors16] and go even further by claiming that we can predict protein abundances from transcript abundances across different human tissues using gene-specific translation factors. However, this claim is probably overstated as rebutted by other authors [Fortelny17], because of the validation procedure at use being highly biased by the high overall gene-wise correlations. On the other hand, Fortelny *et al.* show that tissue-wise (or sample-wise) correlations between transcripts and proteins of a same gene are low, with a median correlation of 0.21. These apparent contradictory results are due to the fact that the variation of protein level across replicates or across biological conditions is in general much lower than the variations of the average mean protein level across different genes (an effect similar to Simpson's paradox [Fortelny17]). This result concurs with the statement of the former review [Liu16], explaining that gene-wise correlations by themselves are not sufficient to infer differences of protein levels between different tissues, replicates or conditions.

Unfortunately, the gene-wise protein/transcript correlation makes poor sense in the case of label-free LC-MS/MS proteomics quantitative data, as the relative abundances are not comparable across proteins (of note, absolute label-based quantification was used in [Edfors16]). However, these results suggest that large transcript variations (as the ones between different genes) should in general produce large protein level variations. Also, the aforementioned distribution of sample-wise correlations [Fortelny17] suggests there are cases where they are significant and meaningful to infer protein levels. Finally, these correlations were computed using Spearman's correlation, which only copes for linear dependency. Non-linear dependency could also be considered, as to reflect the complex underlying biochemical relationship, although it would probably require a larger number

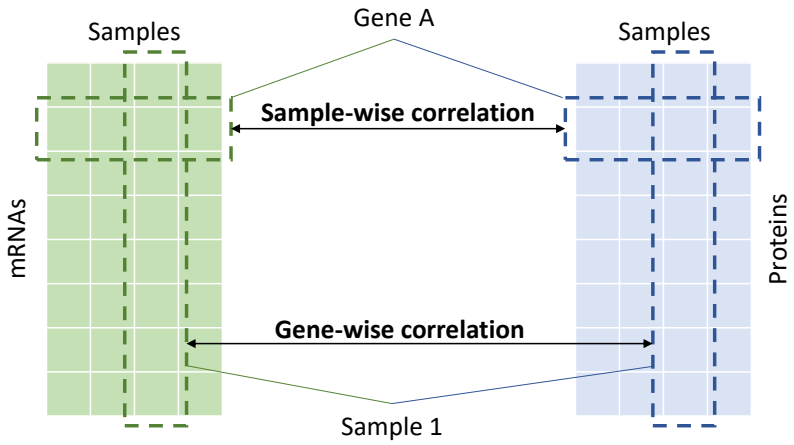

Figure 1.11: Illustration of gene-wise and sample-wise correlations between a transcriptomic and a proteomic dataset. We consider on this figure that proteomics and transcriptomic samples are paired, or at least derive from the same biological condition.

of samples.

1.4.3 Inferring protein abundances with transcriptomic and genomic information

We focus here on tools that can impute or extrapolate protein levels from transcript quantitative information. Note that, in the current state-of-the-art, and to the best of our knowledge, no tool has been specifically used (and benchmarked) to impute proteomics abundances issued by the very technology we focus on in this work, namely LC-MS/MS. Considering the specificities of this instrumental workflows and its consequences on the data distribution, this blind spot of the state-of-the-art is noteworthy.

Some popular tools rely on a common latent representation of different omics modalities. These representations enable to perform numerous multi-omic integration tasks on various omics types, including gene expression arrays. For example, MOFA (Multi-Omics Factor Analysis) [Argelaguet18] relies on the BPCA framework to find a matrix factorization of different omics modalities with common latent factors. In parallel, it includes tools to interpret and use them in various inference tasks. On the other hand, the MixOmics package [Rohart17] relies on multi-blocks Partial Least Squares (PLS) to find a latent decomposition of several omic modalities. For example, with one proteomic and one mRNA dataset, it iteratively identifies pairs of latent factors with maximum covariance. Both MOFA and MixOmics can extrapolate protein levels that have not been measured in samples. MOFA has a multi-omic imputation feature, contrarily to MixOmics. Oppositely, MOFA, and regardless of the analysis type, has never been tested on LC-MS/MS proteomics data.

Whereas these two methods can find common latent structures, regardless of transcript/protein correspondences, the following ones specifically aim at characterizing the conditional distribution of the protein levels with respect to their corresponding transcript levels. Two related work [Nie06, Torres-García09] have tested different approaches (linear Poisson model, gradient boosted trees) on the same dataset for proteomic feature extrapolation. In addition to mRNA levels, they take as input various gene specific features, such as sequence length, protein weight, gene category *etc.* Unfortunately, these works rely on an outdated quantification procedure (based on the count of MS/MS spectra instead of XIC), and the validation method relies on biological criteria specific to the experimental setup. Authors from the same group [Torres-García11] have also proposed gradient boosted trees including various gene-specific features for sample extrapolation in a longitudinal study, but here again, the methodological development and the validation procedure are specific to this experimental design. More recently, Barzine *et al.* [Barzine20] has proposed a deep neural

network to extrapolate protein LC-MS/MS measurements, as well as a cross-species extension. The neural network takes as input the mRNA expression, the identifier of the biological condition, and a binary vector representing gene annotation. Experiments show that including gene annotation information significantly improves performances of the model. Finally, Ochoteco Asension *et al.* [Ochoteco Asensio22] train several linear and non-linear models with a couple of hundreds of features to extrapolate samples. These features are related to mRNA expression, transcript characteristics (strand, length, *etc.*), protein characteristics (length, mass, *etc.*) as well as the expression of micro RNAs and circular RNAs that can have regulatory effects on the transcript considered (cf. "small RNAs" subsection 1.1.1). Surprisingly, their results show that protein specific features such as mass and protein length are among the features with the highest importance in random forest models, along with those related to mRNA expression.

All these works show that including protein or gene characteristics globally helps for extrapolation of samples or proteins, as long as these proteins or related ones (*e.g.*, with similar gene annotations) are seen in the training phase. The fact that the prediction accuracy is significantly increased by including protein-related features (as gene annotation, or molecular characteristics) suggests that these models have learned the feature-protein mapping, which results in imputing by a value close to the observed mean protein's abundance value [Ochoteco Asensio22]. In fact, as discussed in subsection 1.4.2, it is possible to infer the order of magnitude of a protein level knowing mRNA levels and protein characteristics. However, inferring protein variations across biological conditions or replicates, which are often of much smaller amplitude than those with other proteins, remains a difficult task. This difficulty may be the reason why none of these works have proposed a benchmark with other LC-MS/MS proteomics imputation methods, although some of their authors [Barzine20, Ochoteco Asensio22] claimed it could be used in practice for imputation. We hypothesize their performances would not reach those of the state-of-the-art imputation methods for LC-MS/MS data.

In the single-cell community, let us finally note the development of quantitative proteogenomics approaches. In fact, scRNA-seq can now be used in combination with quantification of surface proteins at single-cell resolution. Some methods, notably cTP-Net [Zhou20] and Seurat [Stuart19], have then proposed to impute or extrapolate missing protein levels. For example, cTP-Net [Zhou20] learns the entire joint mapping between scRNA-seq levels in a cell and surface protein levels in the same cell with neural networks. Seurat [Stuart19] is a very general tool that can transfer learning between cells with/without observed protein levels to extrapolate protein levels. Although these single-cell methods leverage enormous sample size (thousands of cells) which is not compatible with bulk LC-MS/MS proteomics and transcriptomics, they are worth mentioning as a source of inspiration.

Overall, we conclude that although there exist methods for general multi-omic imputation and sample or protein extrapolation suited to LC-MS/MS proteomics, so far, no method proposes LC-MS/MS proteomics data imputation that follows a quantitative proteogenomics paradigm, *i.e.*, which leverages transcriptomic data from same or related samples. We have nevertheless identified a preliminary work (*i.e.*, not peer-reviewed yet) presented in a poster at ISMB-EECB 2023 conference ambitioning to explore this path [Gupta23]: the authors rely on Graph Neural Networks to infer missing protein abundances in LC-MS/MS proteomics data, taking as input the abundances of available peptides and mRNAs. Although still preliminary, interesting outputs are expected.

1.4.4 Contributions

We present in chapter 3 our contributions regarding transcriptomic integration for LC-MS/MS proteomics data, which can be summarized as following.

1. **Development of an integrative imputation method for LC-MS/MS proteomics:** In Pirat imputation pipeline, we developed an option to integrate transcriptomic data in a gene

specific manner, naturally compatible with Pirat's model, including the peptide missingness mechanism. This integration can be achieved as long as matching phenotypes are represented in proteomics and transcriptomics data (notably, paired samples are not a necessity, even though they improve the results).

2. **Transcriptomics integration does not deteriorate imputation:** our integration method is based upon observed peptide/transcript correlations, and thus only impacts imputation when these are significant, as experimentally demonstrated.

3. **Increasing coverage of poorly covered proteins:** our approach increases imputation performances on weakly covered proteins (*i.e.*, proteins having only one specific peptide identified and quantified in at least one sample), and in a setting where proteomic and transcriptomic samples are paired.

2

New insights on FDR control in proteomics and applications

This chapter establishes links between different FDR control frameworks, both for differential analysis and peptide identification, gives some perspectives about their usage and presents a practical application of proteomic variable selection in high dimensional setting.

Sommaire

2.1 Motivations

We explained in the previous chapter why FDR control is necessary to LC-MS/MS proteomics. We provided an overall description of the methods currently used, both for peptide identification and differential analysis. An expert reader may have already drawn some parallels between these two applications, although they were developed independently, with more or less robust theoretical foundations. For example, in methods dedicated to peptide identification, the scores of decoy PSMs can be viewed as the distribution of scores under the null hypothesis. Hence, the whole BH and Empirical Bayes frameworks could be applied there, as long as this distribution truly reflects the null hypothesis. In addition, in TDC, a widely used FDR estimator for a given score rejection threshold t (assuming the higher the score, the more confident we are) reads [Levitsky17]:

$$\widehat{\mathrm{FDR}}(t) = \frac{1 + \#\{\text{Decoy PSMs scores} \geq t\}}{\#\{\text{Target PSMs scores} \geq t\}}, \tag{2.1}$$

which clearly resembles the one used in the knockoff procedure (see Equation 1.3).

We bring in this chapter new insights on FDR control in proteomics by investigating these similarities in two original works. In the first one, we have applied and then adapted the knockoff procedure to proteomic differential analysis in a protocol format paper, which illustrates its behavior with several experiments as to draw some parallels with TDC. The second work is a perspective article that details the duality between competition-based approaches (knockoff filters, TDC) and competition-free ones (target-decoy without competition or BH and its Bayesian extensions). In this article, we also rely on the conclusions from the protocol to highlight some weaknesses of TDC (although it is widely used) and propose some improvement paths in view of the theoretically grounded knockoff filters.

Yet, biomarker selection by means of an FDR control is not necessarily the ultimate goal of proteomic analysis, and methodological development of FDR control can only bring us so far. For example, in a clinical context, medical experts often need a score to assess the state of a patient, with known error rates. Although many "FDR-controlled biomarkers" can be used to do so, practitioners anticipate their educated combination should improve the specificity/sensibility ratio. However, doing so is not trivial as FDR control procedures are often univariate. In this context, we also present an applicative work pertaining to the severity score of a disease using several biomarkers which have been individually selected subsequently to an FDR controlled procedure.

2.2 Publication 1: Unveiling the Links Between Peptide Identification and Differential Analysis FDR Controls by Means of a Practical Introduction to Knockoff Filters

2.2.1 Foreword

This first work is a protocol paper of the *Methods in Molecular Biology* (MiMB) series and was thus published as a book chapter –of note, this book has been edited by one of my supervisors. This protocol format is unusual in the biostatistics community: it consists in a detailed step-by-step recipe, code lines included, with many footnotes (for details, interpretations, *etc.*), without conclusion, and that should be accessible for a non statistics or computer science expert. We apply in this work the knockoff procedure to a standard benchmark proteomic dataset. We propose some experiments to compare different variations of the knockoff filters procedure and scoring methods, and interpret their behavior.

This work was conducted at the very beginning of my thesis, and was motivated by various scientific, speculative, and personal aspects. First, the knockoff filters were quite recent at this time, and had not been tested on proteomic data yet, which has the particularity to have few samples for thousands of features. We thus wanted to assess the FDR control quality of this type of dataset. Secondly, and more practically, it gave a short-term objective of writing and results, to maintain motivation in the Covid pandemic period (My PhD contract started during a shutdown, so that my first months of work were remote). Finally, one of my supervisors hoped that a deeper understanding of the knockoff generation framework could give inspiration for a multi-omic imputation method (that could include transcriptomic data), which was one of the major objectives of this PhD thesis. This link was unfortunately too thin, yet, unexpectedly, this work on knockoffs enlightened so many similarities with TDC that we exploited them to draw conclusions from this protocol, and push further the reasoning in the next article.

The reference of this protocol reads:

Etourneau, L.; Varoquaux, N.; Burger, T. Unveiling the Links Between Peptide Identification and Differential Analysis FDR Controls by Means of a Practical Introduction to Knockoff Filters. *Methods Mol. Biol.* **2023**, 2426, 1–24. doi:10.1007/978-1-0716-1967-4_1.

2.2.2 Abstract

In proteomic differential analysis, FDR control is often performed through a multiple test correction (*i.e.*, the adjustment of the original p-values). In this protocol, we apply a recent and alternative method, based on so-called knockoff filters. It shares interesting conceptual similarities with the target-decoy competition procedure, classically used in proteomics for FDR control at peptide identification. To provide practitioners with a unified understanding of FDR control in proteomics, we apply the knockoff procedure on real and simulated quantitative datasets. Leveraging these comparisons, we propose to adapt the knockoff procedure to better fit the specificities of quantitative proteomic data (mainly very few samples). Performances of knockoff procedure are compared with those of the classical Benjamini-Hochberg procedure, hereby shedding a new light on the strengths and weaknesses of target-decoy competition.

2.2.3 Introduction

Controlling the false discovery rate (FDR) is a well-established practice in most -omic approaches, as it answers a pervasive question: Considering quantitative measurements for many covariates (be they genes, transcripts, metabolites, or proteins) in a set of samples split in at least two different biological conditions, how can we shortlist some differentially expressed ones, while controlling the risk of false positives (*i.e.* wrongly selected covariates due to their looking differentially expressed while they are not)? To cope with this, the most commonly used procedure is without a doubt the Benjamini-Hochberg one (BH) [Benjamini95]. However, due to its large field of application, FDR control has focused a lot of additional efforts in biostatistics, and many authors have proposed to improve upon BH FDR control [Benjamini06, Efron72], or have proposed alternative frameworks to do so [Barber15, Candès18, Stephens17].

In the specific case of proteomics, FDR control is not only used in the aforementioned biomarker selection problem. It is also an essential quality control metric when matching experimental fragmentation spectra onto *in silico* spectra (*i.e.*, derived from reference database of protein sequences). However, for historical reasons, the associated FDR control is not performed using classical tools from biostatistics. On the contrary, a rather empirical approach termed target-decoy [Elias07b] is almost exclusively used. It consists in searching two databases: the first one, referred to as target, containing the genuine protein sequences, and another one, referred to as decoy, containing artefactual sequences. Under the assumption that target mismatches and decoy

matches are equally likely, the number of decoy matches can be used to estimate the number of target mismatches, thus opening the door to FDR control.

For a long time, FDR control for peptide identification and for protein differential analysis have been considered as largely independent. However, theoretical connections exist: Notably, it has long been established [Käll08] that if target and decoy databases are searched independently, then the procedure is broadly equivalent to relying on empirical null theory to estimate the FDR in a BH-related way [Efron72]. More recently, it has been shown ([Couté20] that BH procedure could be a user-friendly and computationally attractive alternative to target decoy competition (TDC)[1]. However, recent developments in theoretical biostatistics have made the links between both approaches to FDR control even tighter. Notably, the authors of [Barber15] have proposed to tackle the biomarker research FDR control using an algorithmic procedure akin to that of TDC. Broadly, this novel approach, denoted as "knockoff-filter," works as follows. First, knockoff variables are simulated to be as independent as possible from conditions of samples, but yet preserve the covariance structure of the original variables[2]. Second, a competition is organized between each original variable and its associated knockoff. Third, the proportion of retained knockoffs is used to estimate the proportion of wrongly selected original covariates (see Table 2.1 for a more detailed comparison with TDC). Conversely, authors have recently leverage the theory underlying knockoff filters to propose improved TDC strategies (see [Emery19]).

Overall, the framework of knockoff filters is particularly insightful to provide a global understanding of FDR control in proteomics and the purpose of this protocol is to root such unified view on empirical comparisons using both real and simulated data. Interestingly, the results of these comparisons are compliant with empirical knowledge about the various strengths and weaknesses classically associated to each FDR control method.

2.2.4 Notations

We first start by reviewing commonly used yet conflicting notations in biostatistics and proteomics.

Classical notations in biostatistics

In biostatistics, the false discovery rate (FDR) and the false discovery proportion (FDP) are distinct notions. The FDP corresponds to what was classically and informally referred to as the “true FDR” in proteomics, *i.e.*, the exact proportion of false positives among the proteins that passed the user-defined selection threshold, and therefore deemed as differentially abundant. Of course, except for benchmark artificial or simulated datasets, this quantity is unknown in practice.

The FDR reads as $\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}]$, where $\mathbb{E}$ stands for the expectation, which broadly amounts to the long run average of the FDP on an infinite number of related experiments subject to stochastic fluctuations. This quantity is also unknown but it is much easier to estimate, and such estimate is classically noted $\widehat{\mathrm{FDR}}$. Estimating the FDR is insightful, but unfortunately, not always sufficient [He15]. An unbiased FDR estimate is expected to provide a value closed to $\mathbb{E}[\mathrm{FDP}]$.

[1] *Target-Decoy Competition.* TDC is a specific target-decoy strategy where both databases are concatenated, so that each spectrum can only match to either a decoy or a target spectrum; in other words, both databases are competing for the matches.

[2] *On the generation of knockoffs.* Knockoff variables, under second order approximation, are simulated (section 2.2.4 for mathematical notations) such that: (1) a knockoff variable X_i^{Ko} has the same mean as the original variable X_i; (2) the covariance between knockoff variables is equal to the covariance between the original variables: $cov(X_i^{\mathrm{Ko}}, X_j^{\mathrm{Ko}}) = cov(X_i, X_j)$; (3) the covariance between knockoff variables and original variables is equal to the covariance between the original variables: $cov(X_i^{\mathrm{Ko}}, X_j) = cov(X_i, X_j) \quad \forall i \neq j$; (4) but the variance between a knockoff variable and the original variable is null: $cov(X_i^{\mathrm{Ko}}, X_i) = 0$. Fulfilling all of those constraints in the data simulation process is impossible. Thus, an optimization procedure is used to fulfill them to the best extent possible. Note that knockoff variables are generated without looking at the condition of samples. This ensures that knockoff variables are independent from response y conditionally to original variables, as explained in [Candès18].

Target-Decoy Competition	**Knockoff filter (2nd order approximation)**
1. Construct peptide decoys such that decoy PSMs have same score distribution than erroneous target PSMs	1. For each protein, generate knockoff abundances with same mean and correlation matrices as original abundances.
2. For each real spectrum obtained, find the best match among all targets and decoys, and retain its score.	2. For each protein, compute a score describing whether the original abundances vector or its knockoff best predicts the condition.
3. The number of selected PSMs from decoys at a given cutoff enables to estimate the FDR on selected target PSMs.	3. The number of selected knockoffs at a given cutoff enables to estimate the FDR on selected original proteins deemed differentially abundant.

Table 2.1: Comparison of the target-decoy and knockoff filter procedures for FDR control. (PSM stands for Peptide-Spectrum Match).

However, on a given dataset, this value may be larger or smaller than the FDP. While a slightly too large estimate implies a conservative behavior (there will be less false positives than expected among the shortlisted biomarkers), a too small FDR implies a too liberal quality control and subsequent risks in post-proteomics experiments.

To cope with weaknesses of FDR estimation, FDR control procedures have been developed: they rely on more conservative assumptions that yield slightly lesser selected discoveries at a given cut-off parameter. If we note as $\widehat{\text{FDR}}_\alpha$ the FDR estimate resulting from controlling the FDR at level α (α being classically tuned to 1%) it is likely that

$$\widehat{\text{FDR}}_\alpha \leq \alpha.$$

In other words, if one cuts-off a list of putative biomarkers according to an FDR controlled at 1%, the FDR estimate on this very list is likely to be slightly lower than 1%. However, as the FDP remains unknown, it is the only way to safely assume that the FDP is equal to or lower than 1%.

Classical notations in proteomics

In proteomics, most of the notions described above (section 2.2.4) are conflated. Since the mid-2010s, discriminating between the FDP and the FDR has progressively become standard. However, distinction between FDR (as equal to $\mathbb{E}[\text{FDP}]$), $\widehat{\text{FDR}}$, $\widehat{\text{FDR}}_\alpha$, and α is scarce. The reason is obvious: except for specific methodological publications, most of them are not useful to the community. Indeed, in practice, a proteomic researcher only needs to manipulate α, the cut-off parameter, and to understand that after applying the FDR control accordingly, the FDP is not necessarily strictly equal to α, but possibly slightly smaller. However, the everyday language is error-prone: when one says or writes "We selected the putative biomarkers at an FDR of 1%," what is referred to as FDR is not $\mathbb{E}[\text{FDP}]$, $\widehat{\text{FDR}}$, or $\widehat{\text{FDR}}_\alpha$, but α.

To cope with this, it is possible to rely on other notations. They are not as formal as those of mainstream biostatistics (section 2.2.4) although they are sometimes reported in mathematics works [Bouret18]. However, they are sufficient for a rigorous everyday work in a proteomic lab. Essentially, it amounts to conflate the FDR estimate with α, and to define the FDR control as a procedure which provides the following guarantee with a sufficiently high probability:

$$\text{FDR} \geq \mathbb{E}[\text{FDP}]\,. \tag{2.2}$$

This formulation can be misleading in the sense it gives the impression that the FDR control

procedure indeed controls the FDP[3]. However, it has two advantages: First, it makes the everyday language compliant with the minimum amount of statistical notions possible; second, it simplifies the understanding of other statistical notions such as "q-value" or "adjusted p-value," as using this formalism, they are simply equivalent to the FDR, as detailed in [Burger18]. In the rest of the protocol, the naming conventions resulting from Eq. 2.2 are used, so that FDR refers to α, the FDR level tuned by the practitioner to perform FDR control.

Other notations used in this protocol

Hereafter, the following mathematical notations are used:

1. n: the number of biological samples.
2. p: the number of proteins to include in differential analysis.
3. $X \in \mathbb{R}^{n \times p}$: the matrix of protein abundances, where each row corresponds to a sample and each column corresponds to a protein.
4. X_j: the vector of abundance of the j-th protein, *i.e.* the j-th column of X.
5. $x_{i,j}$: the abundance value of j-th protein for the i-th replicate.
6. y: the vector representing the condition label (numerical value) of biological samples, of length n. For example, the i-th coefficient of y is 1 if the i-th sample comes from the healthy condition, and -1 if it comes from the disease condition.
7. $X^{\text{Ko}} \in \mathbb{R}^{n \times p}$: the knockoff dataset, generated from original dataset matrix X.
8. X_j^{Ko}: the knockoff vector of abundance of the j-th protein.
9. W: the vector of scores of all proteins (only the original ones, not the knockoff), of length p.
10. W_j: the score associated to the j-th protein. A large positive value W_j is evidence that the protein j is differentially expressed. It is typically constructed by comparing the predictive power of X_j and X_j^{Ko} of the sample conditions. Swapping X_j and X_j^{Ko} should swap the sign of W_j. A null W_j means that both X_j^{Ko} and X_j^{Ko} bring the same amount (or lack thereof) of information on the condition.

2.2.5 Material

R version

R version 4.0.3 (or above) is required to use the following packages. We recommend using an integrated development environment like Rstudio to execute the commands of this protocol. It can be downloaded from `https://www.rstudio.com/`.

[3] *FDR or FDP control?* Directly controlling the FDP is more difficult than controlling its expectation, the FDR. However, some papers have tackled this challenge [Romano06, Luo09]. The notion of "control" is tightly defined in statistics and various FDRs have been defined to induce different form of FDR control (for instance exact, strong, or weak controls, such as discussed in [Ge03]). However, Equation 2.2 does not suggests that the procedure controls the FDR (as it now refers to α) but the FDP, although indirectly, through its expectation. We acknowledge this could be a source of confusion, ant tentatively propose to coin the term "indirect FDP control."

2.2.6 Packages

The following packages are necessary:

1. The packages `knockoff`, `lars` ([Hastie13]), and `glmnet` ([Friedman10]) must be installed from the CRAN:

```
install.packages("knockoff")
install.packages("lars")
install.packages("glmnet")
```

2. `cp4p` [Giai Gianetto19] provides two datasets with controlled ground truth: They result from analysis of samples containing different abundance of 48 human proteins spiked in a yeast background [Ramus16]. The p-values from a Welch t-test associated to each protein are also provided, along with functions to apply Benjamini-Hochberg procedure for differential analysis. To install `cp4p` package, it is first necessary to install the BioConductor [Huber15] packages it depends on:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("multtest")
BiocManager::install("limma")
BiocManager::install("qvalue")
```

3. Then `cp4p` can be installed from the CRAN:

```
install.packages("cp4p")
```

4. Finally, load the packages in the environment:

```
library(cp4p)
library(knockoff)
library(lars)
```

Data Format

This protocol relies on a data format which is quite uncommon in proteomics[4]. The input data X on which FDR control is applied should have at least 3 rows, *i.e.* at least biological 3 samples in total are needed. The number of proteins to include in differential analysis can be arbitrary. Values of abundance in X should be $\log_2$-scaled.

For conveniency, we use two datasets in this protocol: A dataset resulting from real mass-spectrometry output, called `LFQRatio25` (section 2.2.6), and a simulated dataset with adjustable parameters (section 2.2.6).

Data loading from cp4p

The following commands enable to load and prepare `LFQRatio25` dataset [Giai Gianetto19]:

1. Load the dataset with the following command:

```
data("LFQRatio25")
```

[4] *Data format.* In proteomics, data tables are generally structured with proteins as rows and replicates in columns. However, `knockoff` and `lars` packages were designed for more general use cases. Hence, they adopt another convention widely used in statistics, with features as columns and samples as rows.

2. Then, abundances values for all 6 samples are extracted to form the rows of the `X_yups` variable:

```
X_yups = t(LFQRatio25[,1:6])
```

3. Similarly, vector `y_yups` contains the condition labels of these samples:

```
y_yups = c(1,1,1,-1,-1,-1)
```

4. For this particular dataset, differentially abundant proteins (or in statistical language, variables under the alternative hypothesis H_1) are known. It is possible to display their name and their index in the list of proteins. These are the 46 first proteins, as the output of this code chunk suggests [5]:

```
mask_human = LFQRatio25$Organism == "human"
names_diff_yups = LFQRatio25$Majority.protein.IDs[mask_human]
idx_diff_yups = which(mask_human)
idx_diff_yups
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17
[18] 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
[35] 35 36 37 38 39 40 41 42 43 44 45 46
```

5. Check the dataset to make sure the same dataset is obtained:

```
head(X_yups[,1:5])
```

```
          [,1]     [,2]     [,3]     [,4]     [,5]
A.R1 31.27392 29.48101 29.80982 29.10410 26.85626
A.R2 31.27147 29.46032 29.84163 29.22384 27.11535
A.R3 31.26327 29.45797 29.83771 29.00945 26.94358
B.R1 29.83022 28.04973 28.41002 27.45505 25.71735
B.R2 29.81413 28.02686 28.38101 27.58463 25.74196
B.R3 29.84867 28.00774 28.42514 27.52028 24.62264
```

Data simulation

The following commands enable to prepare a simulated dataset:

1. The code below randomly generates a dataset broadly akin to `LFQRatio25`. Due to randomness, it will be different from one run to another. To ensure the results are reproducible and to obtain same results as in the remaining of the protocol, use the following optional command to set the random seed[6]:

```
set.seed(1234)
```

[5] *Missing proteins in cp4p.* Th UPS1 mixture contains 48 human proteins. However, according to [Ramus16], only 46 are confidently identified and quantified by mass spectrometry. Therefore, when processing the `LFQRatio25`, only 46 differentially abundant UPS1 proteins are sought.

[6] *Seed in R.* Setting the seed of the random number generator should give you the same sequence of random numbers as presented here. However, different versions of R may yield different sequences of random number due to changes in the pseudo-random number generator.

2. Tune the parameters of the dataset:

```
n_h1 = 50          # Number of proteins differentially abundant
n_rep = 3          # Number of replicates of each condition
p=1500             # Number of proteins
mu = runif(p, 24, 32)
sigma1 = diag(runif(p,0,0.02))
sigma2 = diag(runif(p,0,0.02))
mu_diff = c(runif(n_h1, 0.5, 2)*sign(runif(n_h1, -1, 1)),
    rep(0, p-n_h1))
```

3. Create and concatenate arrays of both conditions:

```
p = length(mu)
X1 = matrix(rnorm(n_rep*p),n_rep) %*% chol(sigma1)
X2 = matrix(rnorm(n_rep*p),n_rep) %*% chol(sigma2)
X1 = t(t(X1)+mu+mu_diff/2)
X2 = t(t(X2)+mu-mu_diff/2)
X_sim = rbind(X1,X2)
y_sim = c(rep(1, n_rep), rep(-1, n_rep))
idx_diff_sim = 1:n_h1
```

4. Check the dataset to make sure there are no mistakes:

```
head(X_sim[,1:5])
```

```
         [,1]     [,2]     [,3]     [,4]     [,5]
[1,] 23.96519 28.55181 27.95301 29.64470 31.05108
[2,] 24.04396 28.21652 27.74679 29.56570 31.51248
[3,] 24.05717 28.39634 27.90406 29.74762 31.56869
[4,] 25.65308 29.55612 29.92890 28.21821 30.47228
[5,] 25.74846 29.63377 29.77653 28.30777 30.32441
[6,] 25.89306 29.58248 29.80624 28.33325 30.52107
```

2.2.7 Methods

This section falls into the following subsections:

1. We explain how to apply the original knockoff-filter procedure to control the FDR for differential expression analysis. Precisely, we show how to (1) generate knockoff variables; (2) compute a score for each protein/knockoff pair; (3) select differentially abundant proteins for a predefined target FDR.

2. We detail how to replace the default scoring strategy with other ones, and compare these alternative knockoff procedures to the classical Benjamini-Hochberg (BH) procedure.

3. We propose some code to illustrate the sensitivity of the knockoff filter procedure to the random generation of knockoffs.

Original knockoff procedure

1. Choose the dataset on which applying the knockoff procedure:

   (a) To apply it on the `LFQRatio25` dataset, use:

```
X_data = X_yups
y_data = y_yups
idx_diff = idx_diff_yups
```

(b) Alternatively, to apply it on the simulated dataset, use:

```
X_data = X_sim
y_data = y_sim
idx_diff = idx_diff_sim
```

For the rest of this section, we will use the `LFQRatio25` dataset.

2. Rescale the data to have null mean and unitary variance for each protein abundance vector (*i.e.* for each X_j)[7]:

```
X_data = scale(X_data)
```

3. Execute these commands to generate the knockoff dataset from original data with a fixed seed[8]:

```
set.seed(1234)
X_data_k = create.second_order(X_data)
```

4. For each protein, compute a score based on the Lasso path of covariates[9]. An inevitable warning concerning the lack of replicates appears: "one multinomial or binomial class has fewer than 8 observations; dangerous ground."

```
set.seed(1234)
W_lasso = stat.lasso_lambdasmax_bin(X_data, X_data_k, y_data)
```

5. Set the value of targeted FDR, compute the resulting threshold, and select proteins for which their score is above this threshold. The `target_fdr` parameter must be a number between 0 and 1. The `offset` parameter determines which FDR estimator to use, it can be set to either 0 or 1[10] . When `offset` is 0, a biased FDR estimate is used, and when `offset` is 1, a

[7] *Data scaling.* This step is particularly important when variable variances span over a large range. To give an order of magnitude, in the `LFQRatio25` dataset, the lowest variance among all covariates equates 0.001 while the largest one equates 9. If the dataset is not scaled, the program used to generate knockoff converges after 10 minutes, while 40 seconds are sufficient with scaled data.

[8] *Warnings in knockoff generation.* We have observed that the following warnings "Reached upper boundary" and "only 0 eigenvalue(s) converged, less than $k = 1$," may appear in some environments. We assume these warnings come from the too high dependence between the columns which corresponds to differentially abundant proteins, yet, the algorithm can still operate.

[9] *Scoring methods.* The `knockoff` package already provides the functions to compute scores according to different methods. A classical scoring method used in the original knockoff procedure is based on the Lasso path of variables, thus we try to apply it first. However, other methods using Lasso are proposed in the package, such as `stat.lasso_lambdadiff_bin` and `stat.lasso_coefdiff_bin` (this last one is not applicable on our data by lack of samples). Also, a method based on random forests is proposed, but during our preliminary experiments, it gave poorer results on our datasets.

[10] *Offset parameter.* The offset parameter corresponds to the difference between the natural FDR estimate and the conservative estimate yielding FDR control. The former reads

$$\frac{\#\text{ of knockoffs selected at } \alpha \text{ level}}{\#\text{ of original variables selected at } \alpha \text{ level}}$$

while the second reads:

$$\frac{\#\text{ of knockoffs selected at } \alpha \text{ level} + 1}{\#\text{ of original variables selected at } \alpha \text{ level}}.$$

This distinction also exists in the TDC procedure [He15] where the natural estimate reads d/t and the conservative ones $(d+1)/t$, where t and d respectively denote the number of selected target and decoy PSMs. In the knockoff literature, the "+1" is termed "offset," and when equal to 0 (respectively 1), it leads to the biased (respectively, non-biased) estimate. The non-biased estimate is then more conservative than the biased estimate.

non-biased, yet more conservative estimate is used.

```
target_fdr = 0.05
thres = knockoff.threshold(W_lasso, fdr=target_fdr, offset=0)
selected_lasso = which(W_lasso >= thres)
```

6. **This step and the following ones are optional, as they can only be applied for a dataset endowed with a ground truth, such as `LFQRatio25` or a simulated dataset.** Display the names of proteins selected as differentially abundant at the FDR tuned with the `target_fdr` parameter (here 0.05).

```
names_diff_yups[selected_lasso]
```

```
[1] P02768upsedyp ALBU_HUMAN_upsedyp - CON__P02768-1
[2] O00762upsedyp UBE2C_HUMAN_upsedyp
[3] P00709upsedyp LALBA_HUMAN_upsedyp
[4] P02788upsedyp TRFL_HUMAN_upsedyp
[5] P06396upsedyp GELS_HUMAN_upsedyp
[6] P12081upsedyp SYHC_HUMAN_upsedyp
```

7. This code instantiates useful functions to compute the FDP and power from ground truth data. For a certain selection level α, the power is defined as

$$\text{Power}_\alpha = \frac{\#\text{ of selected original variables under } H_1}{\#\text{ of original variables under } H_1}.$$

Where H_1 denotes the alternative hypothesis, *i.e.* "the protein is differentially abundant." The power gives a measure of how well our selection covers all the proteins differentially expressed:

```
compute_fdp = function(selected, nonzero) {
    if (length(selected) != 0) {
        return(1-sum(nonzero %in% selected)/length(selected))
    }
    return(0)
}

compute_power=function(selected, nonzero) {
    if (length(selected) != 0) {
        return(sum(nonzero %in% selected)/length(nonzero))
    }
    return(0)
}
```

8. The following code computes the FDP and power of the procedure for a user-defined range of target FDRs (for both offset values):

```
FDR = seq(0,0.5,0.04)
template = rep(0,length(FDR))
FDP = list(template, template)
POWER = list(template, template)

for (t in 1:length(FDR)) {
  for (offs in 1:2) {
    thres = knockoff.threshold(W_lasso, fdr=FDR[t],
        offset=offs-1)
    selected = which(W_lasso >= thres)
```

```
    FDP[[offs]][t] = compute_fdp(selected, idx_diff)
    POWER[[offs]][t] = compute_power(selected, idx_diff)
    }
}
```

9. Using the results computed at the previous step, the following code displays the FDP and power as a function of the FDR (see Figure 2.1 for `LFQRatio25` and Figure 2.2 for simulated dataset):

```
par(pty='s')
cols = c("red","blue", "black")
plot(FDR, FDR, type='l', ylab = "FDP", xlab = "FDR",
    ylim=c(0,0.5), xlim=c(0,0.5))
lines(FDR, FDP[[1]], col="red")
lines(FDR, FDP[[2]], col="blue")
points(FDR, FDP[[1]], col="red", pch=1)
points(FDR, FDP[[2]], col="blue", pch=2)
legend("topleft", legend=c(0,1, "y=x"), col=cols,
    pch=c(1,2,-1), lty = 1, title="Offset")


plot(1, type="n", ylab = "Power", xlab = "FDR",
    ylim=c(0,0.4), xlim=c(0,0.5))
lines(FDR, POWER[[1]], col="red")
lines(FDR, POWER[[2]], col="blue")
points(FDR, POWER[[1]], col="red", pch=1)
points(FDR, POWER[[2]], col="blue", pch=2)
legend("topleft", legend=c(0,1), col=cols, pch=c(1,2),
    lty = 1, title="Offset")
```

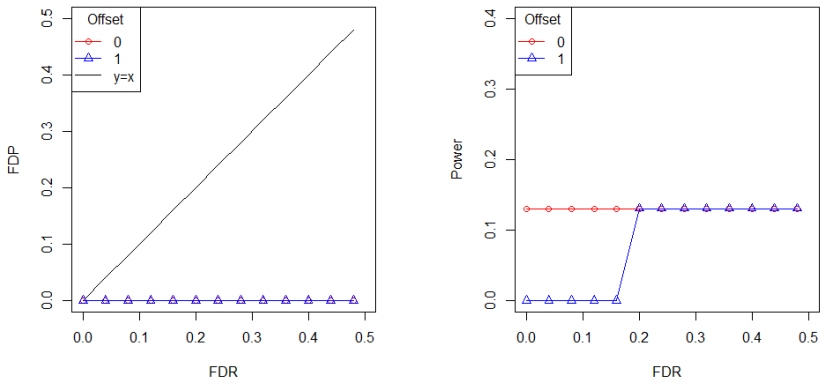

Figure 2.1: FDP and power vs. *FDR for `LFQRatio25` dataset, with and without offset, for the knockoff filter procedure with Lasso-based scores.*

We notice that FDP and power curves on Figure 2.1 and 2.2 are almost always horizontal. This means that variables selected remain the same whatever the FDR target chosen. When the offset equates 1 (unbiased estimator), no proteins are deemed differentially expressed below a certain value of FDR. Thus, even though their are no false positive, there are no true positive either, making the FDR control through knockoff filters practically useless.

We mainly explain this over-conservativeness by the usage of variable selection with the Lasso algorithm, at the step of W scores computation. In fact, in the setting $n << p$, the Lasso algorithm

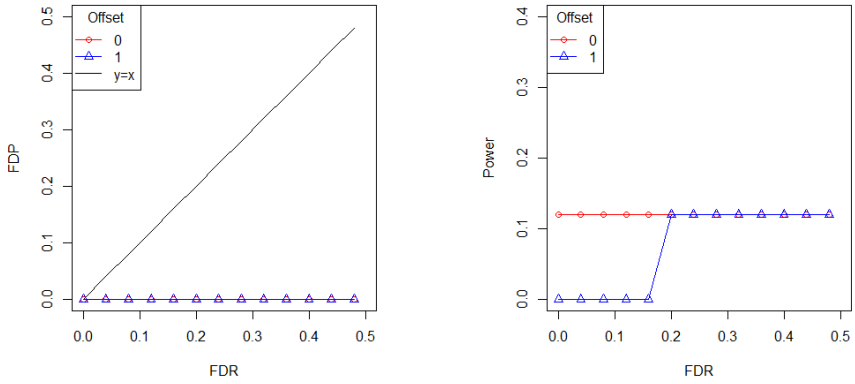

Figure 2.2: FDP and Power vs. *target FDR for the simulated dataset, with and without offset, for knockoff procedure with Lasso-based scores.*

will only select n variables. This is problematic for differential expression analysis where the total number of samples rarely exceeds the number of *a priori* differentially expressed proteins. On top of that, as very few covariates are selected, and some original variables are much more differentially abundant than all the others, knockoff variables are almost never selected. Thus, estimating the number of false discoveries from the number of selected knockoffs is not appropriate in our cases. These efficiency of variable selection with Lasso is thoroughly discussed in [Zou05].

Scoring methods based on forward stagewise regression and *t*-test

Preliminary experimental comparisons highlighted the knockoff procedure accuracy highly depends on the chosen feature selection algorithm. We herefater describe two procedures that we found to address the issue described above (section 2.2.7). The first scoring method consists in using forward stagewise selection (FS) algorithm[11]. The second one is derived from the variable selection procedure classically used in proteomics: it amounts to computing a t-test p-value for both original and knockoff variables; then, the final score (*i.e.*, W_j) is defined by the log difference of p-values (LDP) obtained between each original variable and its knockoff.

1. To instantiate the functions that compute the W_i's for the FS and LDP methods, use the following chunks of code (it is advised to run them both, so as to allow subsequent comparisons):

   (a) For the FS method:

```
stat_forward_sel=function(X, X_k, y) {
    Xconcat = cbind(X, X_k)
    res = lars(Xconcat, c(1,1,1,-1,-1,-1), type="for",
        use.Gram = FALSE)
    lambdas = rep(0, 2*ncol(X))
    lambdas[res$entry != 0] = res$lambda[res$entry]
    W_fs = lambdas[1:ncol(X)]-lambdas[-(1:ncol(X))]
```

[11] *Lasso vs forward stagewise.* From a theoretical point of view, the forward stagewise algorithm behaves very much like the Lasso [Efron04]. However, it copes with the issue of the Lasso being unable to select more variables than the number of samples, an essential feature in proteomics. The FS score we use is based on the norm of predictors coefficients when a given variable enters the model, similarly to the Lasso method. This method is already proposed in the `knockoff` package and the associated paper [Candès18] through the `stat.forward_selection` function. However, we use slightly modified approach relying on the `lars` package to obtain an equivalent regularization term, instead of variable selection ranking.

```
    W_fs
}
W_fs = stat_forward_sel(X_data, X_data_k, y_data)
```

(b) For the LDP method:

```
stat_log_diff_pval=function(X, X_k) {
    Xconcat = cbind(X, X_k)
    pvals = apply(Xconcat, 2, function(x){res =
        t.test(x[1:3], x[4:6]);return(res$p.value)})
    pvals_or = pvals[1:(length(pvals)/2)]
    pvals_k = pvals[(length(pvals)/2+1):length(pvals)]
    W_pvals = (-log(pvals_or)+log(pvals_k))
    W_pvals
}

W_ldp = stat_log_diff_pval(X_data, X_data_k)
```

2. Plot the histogram of W_i's to better visualize the selection process (see Figure 2.3 for `LFQRatio25` dataset):

```
hist(W_ldp[W_ldp!=0], col=c(rep("red", 2), rep("grey", 4),
    rep("blue", 11)), main="Histogram of W", xlab="W")
axis(1, at=c(-5, -2, 0, 2, 5, 10))
```

3. To illustrate the interest of using FS and LDP within the knockoff filter procedure, we compare those two approaches with the classically used Benjamini-Hochberg (BH) procedure. Depending on the dataset being `LFQRatio25` or the simulated one, the code differs:

(a) With `LFQRatio25`, the p-values resulting from Welch t-test are provided in the dataset:

```
pvals = LFQRatio25[,7]
res = adjust.p(pvals, pi0.method = 1)
```

(b) With the simulated dataset, p-values must be computed beforehand (a Welch t-test is also used here):

```
pvals = apply(X_data, 2, function(x){res = t.test(
    x[1:n_rep], x[(n_rep+1):(2*n_rep)]);
    return(res$p.value)})
res = adjust.p(pvals, pi0.method = 1)
```

4. Compute the FDP and power for BH and knockoff filter procedure with LDP and FS methods (with offset=1), at different FDR levels:

```
FDP = list(template, template, template)
POWER = list(template, template, template)
W_list = list(W_fs, W_ldp)

for (t in 1:length(FDR)) {
  for (W_idx in 1:2) {
    thres = knockoff.threshold(W_list[[W_idx]], fdr=FDR[t],
        offset=1)
    selected = which(W_list[[W_idx]] >= thres)
    FDP[[W_idx]][t] = compute_fdp(selected, idx_diff)
    POWER[[W_idx]][t] = compute_power(selected, idx_diff)
  }
  selected_bh = which(res$adjp$adjusted.p<=FDR[t])
  FDP[[3]][t] = compute_fdp(selected_bh, idx_diff)
```

```
  POWER[[3]][t] = compute_power(selected_bh, idx_diff)
}
```

5. Finally plot the FDP and power *vs.* FDR level, as illustrated on Figure 2.4 and 2.5, respectively for the `LFQRatio25` and simulated datasets):

```
par(pty='s')
cols = c("red", "blue", "orange")
leg = c("Knockoff w F.S.", "Knockoff w log diff.", "B-H.")
plot(FDR, FDR, type='l', ylab = "FDP", xlab = "FDR",
    ylim=c(0,0.6), xlim=c(0,0.15))
for (i in 1:3) {
  lines(FDR, FDP[[i]], col=cols[i])
  points(FDR, FDP[[i]], col=cols[i], pch=i)
}
legend("topleft", legend=leg, col=cols, pch=1:3,
    title="Procedure")

plot(1, type="n", ylab = "Power", xlab = "FDR",
    ylim=c(0,1.2), xlim=c(0,0.15))
for (i in 1:3) {
  lines(FDR, POWER[[i]], col=cols[i])
  points(FDR, POWER[[i]], col=cols[i], pch=i)
}
legend("topleft", legend=leg, col=cols, pch=1:2,
    title="Procedure")
```

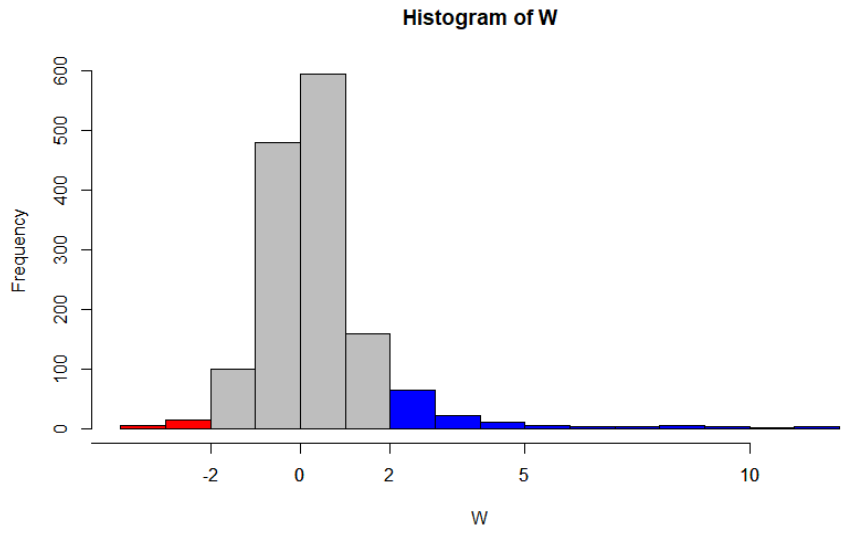

Figure 2.3: Histogram of scores W_i's obtained with log diff of p-values scoring method, on `LFQRatio25` dataset. The blue area correspond to original variables that are selected, and the red area represent knockoff variables selected, both at a threshold of 2 (hence, a conservative FDR estimate at a selection threshold of 2 reads $\widehat{FDR} = \frac{\text{red area}+1}{\text{blue area}}$).

We observe that the knockoff filter procedure with LDP broadly follows the same trend as the BH one on `LFQRatio25` (see Figure 2.4). By construction, the LDP scores is never null, yielding a rather symmetric distribution (see Figure 2.3). The largest positive scores (depicted in the right hand tail) result from differentially abundant proteins, while the left hand one amounts to selected knockoff proteins. The distribution being more symmetric than when using the Lasso, it is possible

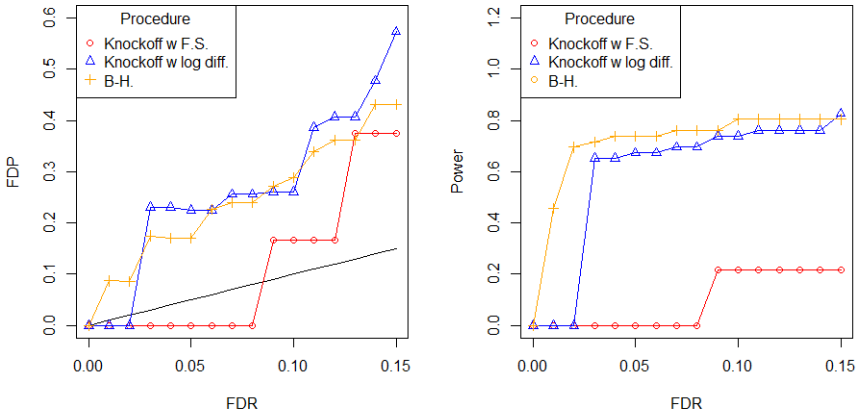

Figure 2.4: FDP and power vs. *target FDR for knockoff filter procedure with offset=1 applied with forward stagewise selection and log diff of p-values scoring, and Benjamini-Hochberg procedure, obtained with* `LFQRatio25`.

to select a larger subset of proteins at a given FDR. However, when using the FS based scores, knockoff filters roughly behaves as with the Lasso, yielding a greater but yet insufficient power.

Finally, the BH procedure also yields anti-conservative results on `LFQRatio25`, as the FDP is always higher than the FDR. However, this can be explained by other preprocessing steps (match between runs, normalization, imputation, *etc.*) which tends to shrink the within-condition variance prior to differential analysis as well as to increase the risk of false positives that are not accounted by FDR control. Indeed, Benjamini-Hochberg is conservative on simulated data (see Figure 2.5).

Sensitivity of FDR control to knockoff used

Knockoff generation with `create.second_order` function (section 2.2.7) involves the random draw of a knockoff matrix (similarly to the random generation of decoy sequences). Hence, on a given dataset, running two consecutive FDR control procedures with knockoff filters should lead to slightly different results. We hereafter propose several experiments to illustrate the sensitivity of the knockoff filter procedure to the knockoff generation, as well as to evaluate its magnitude.

1. Generate 30 knockoff datasets and store them in a list (depending on the machine, this step may last between 30 minutes to an hour):

```
set.seed(3456)
n_k = 30
l_k = list()
for (i in 1:n_k) {
  l_k[[i]] = create.second_order(X_data)
}
```

2. Apply the knockoff filter procedure to each knockoff series, with FDR varying from 1% to 15%. In this example, the scoring method used is LDP. For all the knockoff series, the effective FDP *vs.* FDR curves are iteratively plotted, leading to a display akin to that of Figure 2.6. The proteins selected at an FDR of 5% for each knockoff series are retained in a matrix referred to as `scores`:

```
par(pty='s')
FDR = seq(0.01,0.15,0.01)
FDP <- POWER <- matrix(rep(0,n_k*length(FDR)), nrow=n_k)
```

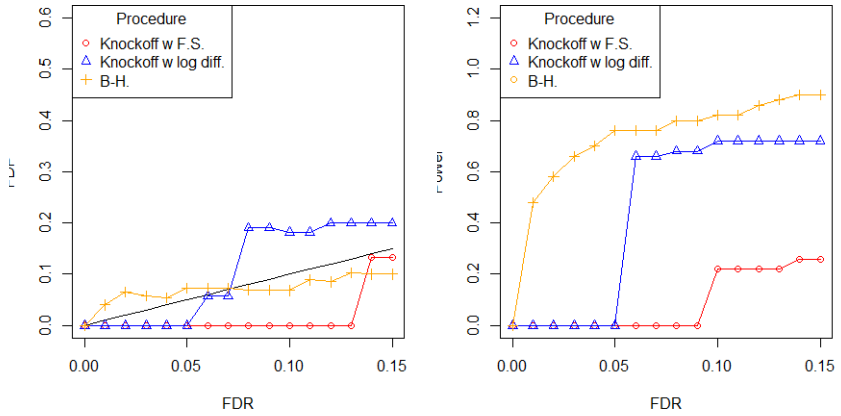

Figure 2.5: FDP and power vs. *target FDR for knockoff procedure with offset=1 applied with forward stagewise selection and log diff of p-values scoring, and Benjamini-Hochberg procedure, obtained with simulated data.*

```
scores = matrix(rep(0, n_k*ncol(X_sim)), nrow=n_k)

plot(FDR, FDR, type='l', ylab = "FDP", xlab = "FDR",
    ylim=c(0,0.7), xlim=c(0,0.15))
for (i in 1:n_k) {
  W = stat_log_diff_pval(X_data, l_k[[i]])
  for (t in 1:length(FDR)) {
    thres = knockoff.threshold(W, fdr=FDR[t], offset=1)
    selected = which(W >= thres)
    FDP[i,t] = compute_fdp(selected, idx_diff)
    if (FDR[t] == 0.05) {
      scores[i,selected] = 1
    }
  }
  lines(FDR, FDP[i,], col=i)
}
legend("topleft", legend = "y=x", lty=1, col="black")
```

3. Finally, plot a heatmap featuring the `scores` matrix which highlights with different colors the selected proteins under H_0 and H_1 for each knockoff filter series, at an FDR target of 5%. (see Figure 2.7):

```
par(mar=c(5, 5, 2, 8), xpd=TRUE, mgp=c(1,1,0))
heights = sort(colSums(scores), decreasing = T,
    index.return = T)
heights_in_plot = heights$ix[heights$x>0]
submat = scores[,heights_in_plot]
submat[(submat == 1) & t(matrix(rep(heights_in_plot>46,
    nrow(scores)), ncol=nrow(scores)))] = 2

image(t(submat), col=c("grey", "blue", "red"),
    xlab="Proteins (selected at least once)", axes=F)
mtext(text=c(paste("Knockoff",c(1,15,30))), side=2, line=0.1,
    at=seq(0.0,1,1/2), las=1, cex=0.9)
legend("topright",inset=c(-0.23, 0),
    legend=c("Selected H_0","Selected H_1", "Not selected"),
```

```
    fill=c("red","cyan", "grey"))
```

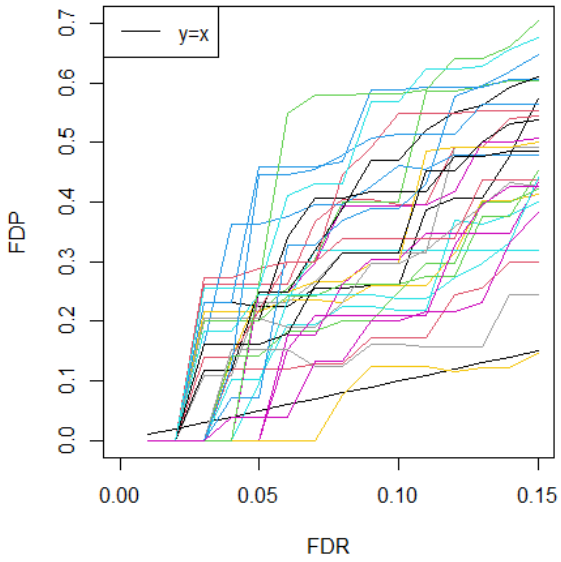

Figure 2.6: Curves of FDP vs. *FDR for 30 different Knockoff procedure, applied with log diff of -values score on* `LFQRatio25` *dataset.*

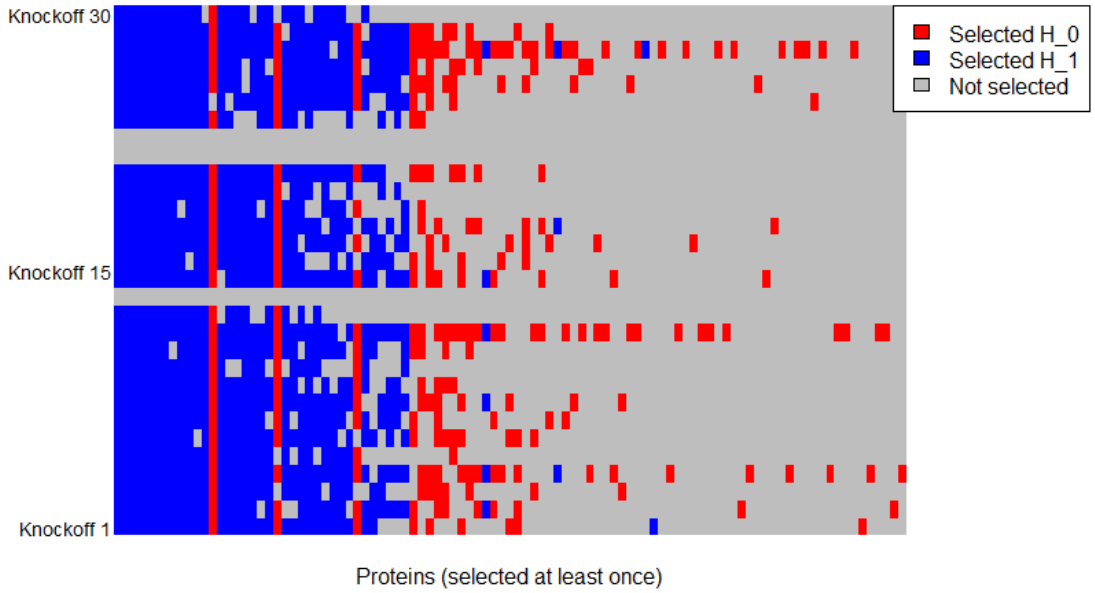

Figure 2.7: Proteins selected according to 30 different knockoff procedure iterations (using LDP score) on the `LFQRatio25` *dataset. Blue cells depict original differentially abundant proteins (human proteins) that were selected using a given knockoff. Similarly, red cells depict non-differentially abundant proteins (yeast proteins) mistakenly selected. Proteins are sorted from the most selected one (left hand side) to the least selected one (right hand side).*

Figure 2.5 and 2.7 emphasize the important variability resulting from the random nature of knockoff filters. To counter this variability, [Nguyen20] proposes a method to aggregate multiple knockoffs. In fact, similar sensitivity has already been commented upon with target-decoy procedures [Keich19], so it seems to be a problem ubiquitous to FDR control procedures which involve simulating artifactual data under the null hypothesis. Finally these observations provide an intuitive support for the tools described in [Emery19], which relies on multiple decoy databases to construct a knockoff-like score.

2.3 Publication 2: Challenging Targets or Describing Mismatches? A Comment on Common Decoy Distribution by Madej et al.

2.3.1 Foreword

As a follow-up to the MiMB protocol, we give a global view on FDR control at peptide identification step in a perspective article published in *Journal of Proteome Research*. To submit it, we took the opportunity of commenting an article by Madej *et al.* [Madej22] that proposed an FDR control method mixing both competition based and free approaches, which naturally addressed an issue regarding their theoretical foundations and assumptions.

Concretely, we first give a recap of FDR control methods at peptide identification and explain the differences between the competition based and free approaches. We then push the parallel between TDC and knockoff procedure further than in the book chapter and highlight discrepancies between them two. Finally, we set warnings inherent to the competition procedure, which extend those formulated in the previous work, and give some perspectives on how TDC could benefit from more theoretical foundations.

This article is referenced as:

Etourneau, L.; Burger, T. Challenging Targets or Describing Mismatches? A Comment on Common Decoy Distribution by Madej et Al. *J. Proteome Res.* **2022**, 21 (12), 2840–2845. https://doi.org/10.1021/acs.jproteome.2c00279.

2.3.2 Abstract

In their recent article, Madej *et al.* (Madej, D.; Wu, L.; Lam, H. Common Decoy Distributions Simplify False Discovery Rate Estimation in Shotgun Proteomics. *J. Proteome Res.2022, 21* (2), 339–348) proposed an original way to solve the recurrent issue of controlling for the false discovery rate (FDR) in peptide-spectrum-match (PSM) validation. Briefly, they proposed to derive a single precise distribution of decoy matches termed the Common Decoy Distribution (CDD) and to use it to control for FDR during a target-only search. Conceptually, this approach is appealing as it takes the best of two worlds, *i.e.*, decoy-based approaches (which leverage a large-scale collection of empirical mismatches) and decoy-free approaches (which are not subject to the randomness of decoy generation while sparing an additional database search). Interestingly, CDD also corresponds to a middle-of-the-road approach in statistics with respect to the two main families of FDR control procedures: Although historically based on estimating the false-positive distribution, FDR control has recently been demonstrated to be possible thanks to competition between the original variables (in proteomics, target sequences) and their fictional counterparts (in proteomics, decoys). Discriminating between these two theoretical trends is of prime importance for computational proteomics. In addition to highlighting why proteomics was a source of inspiration for theoretical biostatistics, it provides practical insights into the improvements that can be made to FDR control methods used in proteomics, including CDD.

2.3.3 A short history of FDR in biostatistics and in proteomics

A False Discovery Rate (FDR) is a statistical estimate of the expected proportion of features that pass by chance a significance threshold (a.k.a., false discoveries). With the advent of high-throughput analyses, the number of measurable features have sky-rocketed. To avoid producing a proportional increase in false discoveries, it has become essential to control for the FDR (*i.e.*, to conservatively select features based on the FDR). Although the starting point of FDR theory is unquestionably dated to 1995 with the publication of the seminal article by Benjamini and Hochberg (BH) [Benjamini95], few later publications acknowledge the importance of pre-existing work [Benjamini10, Goeman14]. After few technical improvements [Yekutieli99, Benjamini16] between 1995 and 2000, the subject really gained momentum with the publication of the human genome [Venter01], which revealed how high-throughput biology could dramatically take advantage of these hitherto purely theoretical advances. The early 2000s thus saw the emergence of several innovations. On the theoretical side, a group of researchers from Stanford reformulated the BH framework to better fit applications in biostatistics [Storey02, Storey03, Efron02, Efron01]. This notably led to the now well-established concepts of q-value (or adjusted p-value) and empirical null estimation.

Meanwhile, in the proteomics community, questions akin to FDR estimation showed up under several names (*e.g.*, “false identification error rates” [Keller02] in 2002 and "false-positive identification rate" [Masselon03] in 2003). It also coincided with the moment when Elias and Gygi *et al.* [Peng03] formulated their intuitions about false positive simulation through decoy permutations, preceding what is now known as Target-Decoy Competition [Elias07a] (TDC). It should be noted that this was a complete conceptual breakthrough at the time, as there was no statistical theory to support the idea that fictional variables (*i.e.*, decoy sequences) created from the original ones (*i.e.*, target sequences) could be used to control for FDR. This is also why decoy databases were soon proposed for use in ways that were more compliant with the pre-existing theory of FDR control. Notably, in 2007–2008, two groups independently proposed that target and decoy searches be performed on separate databases, *i.e.*, without organizing a competition between them (hereafter referred to as TDwoC, to emphasize the absence of competition) [Käll07, Martínez-Bartolomé08]. The first article [Käll07], shorter and more conceptual, became the benchmark (despite the fact that the estimator proposed was far from optimal [Keich15]). This article notably established the theoretical exactness of TDwoC by linking the approach to empirical null estimation [Efron02, Storey03], a concept to which one of the coauthors had also contributed. In addition, they raised concerns about TDC in the conclusions of the article, as in their opinion, the additional competition procedure made it difficult to derive the distribution of target mismatch scores (a.k.a. target null PSMs). However, despite this warning as well as rare voices pointing out the apparent inaccuracy of TDC [Cooper12], the TDC approach progressively became the reference method over the course of the following decade.

This gap between practical approaches to FDR in proteomics and theoretical background in biostatistics was tentatively filled by He *et al.* (in works that remained largely unpublished [He15, He18a] until recently [He18b]). Briefly, these authors demonstrated that FDR could be controlled (at the peptide-only level, as opposed to the more classically considered PSM level) using decoy sequences. They connected their demonstration to simultaneously emerging studies from the Candès group [Barber15]. Although Barber and Candès’ seminal work unleashed an important and ongoing renewal of FDR theory in the statistics community [Candès18, Gimenez18, Ge21, Xing21], it may seem old to proteomics researchers, as its core idea is to fabricate uninteresting putative biomarkers in silico (*i.e.*, fictional variables referred to as “knockoffs”) and to use them to challenge each real putative biomarker through pairwise competition.

2.3.4 Two distinct approaches to FDR

Today, the FDR can be controlled in two ways, both in theoretical statistics and in proteomics: based on a description of how false-positives distribute; or based on competition, by challenging the variables of interest with fictional ones (a.k.a. knockoffs or decoys). We hereafter summarize the two trends, along with their specificities.

"Describing mismatches" or the null-based approach

The oldest approach is based on a simple rationale: The scores of observations we are not interested in (spectrum/peptide mismatches) form the so-called null distribution in statistics. If enough is known about the null distribution, then it is possible to "subtract" it from the distribution observed. We will be left with observations that lie beyond the null distribution, which can therefore be considered of significant interest (discoveries); in sum, to be correct PSMs. Despite a complex mathematical vehicle, necessary for statistical guarantees, the original BH procedure is the first and simplest implementation of this approach. However, this procedure relies on a strong assumption: that the scores distributed are p-values, as the null distribution of such values only is known to be uniform [Burger18], at least in theory [Giai Gianetto16, Wieczorek19]. As such, the BH procedure is the natural tool to control for the FDR when analyzing differential expression, where statistical tests are applied to all putative biomarkers. However, it can also be applied for peptide identification, provided PSM scores can be converted into p-values [Couté20, Fancello22].

If no p-value can be determined from the PSM scores, the approach remains valid, but an additional preliminary step is necessary. The purpose of this step is to estimate how PSM scores distribute under the null hypothesis (to keep the subsequent subtraction from the observed distribution feasible). This extension of the BH framework is naturally referred to as "empirical null estimation" (or "Empirical Bayes estimation" when the alternative hypothesis is also accounted for). Related approaches have been used in proteomics for two decades [Keller02, Choi08], and are still under investigation [Prieto20]. TDwoC is their quintessence, as it provides a universal, conceptually simple, and easy-to-implement means to derive the distribution of random matches.

To summarize, when decoy sequences are used for empirical null modeling, they must be considered as a whole, essentially as a means to describe the data under the null hypothesis. As this distribution will subsequently be subtracted from the target distribution, it must remain unaltered. Notably, this implies that all the decoys must be accounted for, which seems incompatible with selecting only a subset of them based on how they compete against target sequences.

"Challenging targets" or the competition-based approach

The second approach makes no attempt to elicit the null distribution. It only assumes the existence of a procedure to mimic the features of interest (be they amino acid sequences or quantitative vectors describing hypothetical differential abundances). Each feature thus fictionalized is used to challenge an original feature through pairwise competition. Then, the FDR can be estimated thanks to the overall analysis of all the competition results. Despite regular use in proteomics over 15 years as part of TDC, this approach has only recently been theoretically justified [He15, Barber15]. Therefore, it is now tempting to consider it as the posthoc support for TDC applications that has been sought for more than a decade. Unfortunately, detailed analysis of this so-called knockoff framework, that we will summarize below, appears to partially contradict this view.

Knockoff-based FDR is notably applied in biomarker discovery. It considers as input an array-like quantitative data set, where each covariate (*e.g.*, a vector of abundances, polymorphism presence/absence across samples, etc.) is a potential biomarker. The first step of the method aims to fabricate a "knockoff" array of the exact same size as the original one. This knockoff array can be pictured as similar to the permutated array used in permutation-based FDR [Tusher01], except for two differences: first, the knockoff values are randomly generated from scratch rather than by

shuffling the samples. Second, whereas permutation-based FDR relies on an overall description of the null distribution (like BH), knockoff covariates are paired with the originals in order to fit the so-called *exchangeability property* [Barber15]. Respecting this property is challenging, notably when the number of features considerably exceeds the number of samples [Candès18]. Multiple methods have now been described to generate the knockoffs. Figure 2.8 presents the original procedure, which, despite its complexity, only approximately obeys the exchangeability property. More specifically, the covariance matrix for the joint original and knockoff covariates must read as:

$$\mathrm{cov}(X,\tilde{X}) = \begin{pmatrix} \Sigma & \Sigma - \mathrm{diag}(s) \\ \Sigma - \mathrm{diag}(s) & \Sigma \end{pmatrix} \tag{2.3}$$

where X and $\tilde{X}$ refer to original and knockoff covariates, respectively, where Σ is the estimated covariance matrix of X, and $\mathrm{diag}(s)$ is a diagonal matrix build upon a vector $s > 0$. To enable sampling of the knockoff covariates, s must be chosen so that $\mathrm{cov}(X,\tilde{X})$ is invertible. In addition, the power of the FDR control procedure depends directly on the coefficients of s being large enough. As a result of this trade-off, tuning s in a high-dimensional setting is challenging. In our view, this clearly illustrates the difficulty of fabricating sufficiently realistic fictional variables. From a more practical viewpoint, TDC-based FDR presents a similar pitfall. As formally defined by He *et al.* [He15], the accuracy of TDC depends on whether the decoy database complies with the *Equal Chance Assumption* during the subsequent competition. Although it is necessary to assume equally probable target mismatches and decoy matches to control the FDR at the peptide level, little is known about the true validity of the assumption. Notably, it has already been reported that instrument mass tolerance filters [Couté20] or on gene expression filters [Fancello22] can affect the correctness of this assumption. Therefore, many other experimental details may have similar impacts. Likewise, it has recently been reported that the diversity of decoy fabrication methods (see next section) or competition modes [Lin21, Lin22] affects the liberal/conservative behavior of the FDR estimate, and we postulate that the knockoff framework may be useful in explaining these effects.

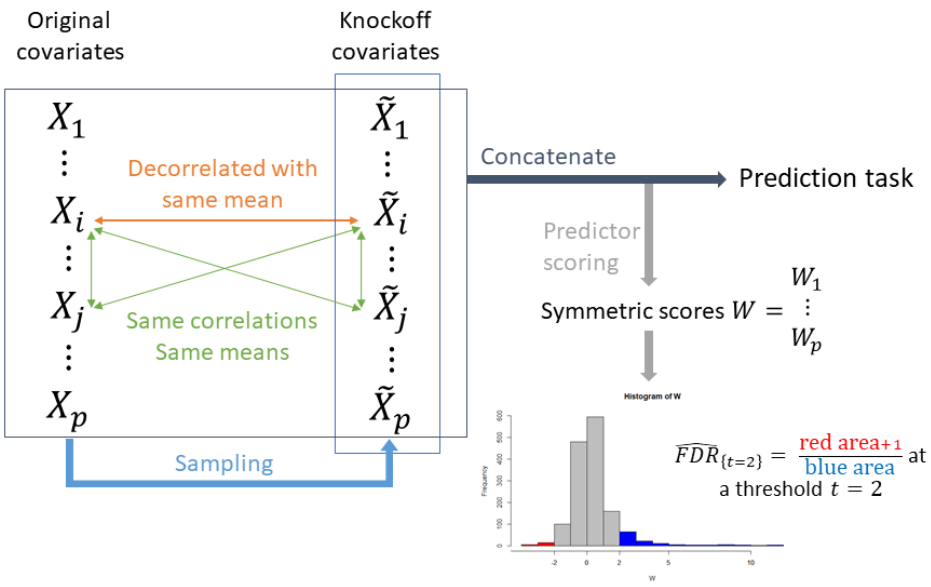

Figure 2.8: General framework of FDR control with knockoff variables (p denotes the number of covariates).

After the knockoffs were generated, a competition is organized between each pair of original

and knockoff covariates. This competition produces a symmetric score W_i measuring which of the i-th original/knockoff covariates is the best suited to a dedicated task (for differential analysis, this task is classically predicting biological status). The more negative the score, the more it indicates the knockoff outperformed the original feature; and conversely, the more positive the score, the more it indicates the original feature outperformed the knockoff (zero corresponds to a draw between the two). Finally, FDR control at level α is achieved as follows: the minimum score threshold t is chosen such that the ratio $(K_t+1)/O_t$ is lower than α, where K_t (respectively O_t) is the number of original/knockoff pairs with relative scores lower than $-t$ (respectively larger than t), see Figure 2.8. All original covariates with a relative score above t are selected. At this point, the parallel with TDC is easy to draw. First, as argued by Keich *et al.* [Keich19], searching a single concatenated database (the original TDC formulation) or two separate databases followed by a competition step is equivalent. Thus, TDC implements a pairwise competition between the target and decoy sequences best matching each spectrum. Second, the FDR control formula is the same as that recently adopted for peptide identification [He15, Keich19, Levitsky17] (as opposed to the original formula [Peng03, Elias07a]). However, at this stage, a small difference emerges. According to the knockoff procedure, the TDC scores (*i.e.*, the W_i's) should be obtained using an antisymmetric function (meaning that $f(x,y) = -f(y,x)$). In practice, the following formula is used:

$$\text{Score}(PSM_i) = W_i = f(Z_i^t, Z_i^d) = \text{sign}(Z_i^t - Z_i^d) \times \max(Z_i^t, Z_i^d) \tag{2.4}$$

where PSM_i is the PSM associated to the i-th spectrum, and Z_i^t and Z_i^d refer, respectively, to the best target and decoy scores with respect to this spectrum. Using this formula, the score complies with the knockoff framework. However, whereas one should retain a null score when $Z_i^t = Z_i^d$ in classical TDC implementations, Z_i^t is often returned instead, thereby breaking the antisymmetry property. Although justified by practical considerations (if a random amino acid sequence appears to be equivalent to an existing peptide, it should not prevent the peptide from being identified), it may hamper the overall statistical correctness of the procedure in practice.

To summarize, although the parallels between TDC and knockoffs are strong, three discrepancies should be kept in mind: the statistical framework cannot handle FDR control at PSM level; the TDC score may not be perfectly symmetrical; and the procedure used to shuffle amino acids does not guarantee compliance with the exchangeability property.

To conclude on the competition-based approach, introducing a pairwise competition step between the fictional and the original variables corresponds to a significant change in the FDR control procedure: it is supported by a distinct mathematical theory and requires specific implementations to work. This calls for caution in the proteomics community, where switching between frameworks has essentially been taken to mean the target and decoy databases are concatenated (competition-based approach) or kept separate (null-based approach).

2.3.5 Perspectives inspired by this history

As we have seen, FDR control is possible using two distinct and orthogonal mathematical theories. When applied to peptide identification, both can rely advantageously on decoy generation. However, the role of the decoy database fundamentally differs, depending on whether the method applied relies on BH and its empirical null extensions (then, decoys are used to "describe the mismatches") or whether it relies on a competition-based approach (then, decoys must adequately "challenge the target sequences"). In our view, this fundamental distinction is worth highlighting, as it provides interesting cues to analyze and improve the tools used for peptide identification.

First, TDC improvements can be expected if we better acknowledge the requirements for generation of knockoffs. Notably, the discrepancy between the TDC scoring function and knockoff requirements (see above) should encourage the investigation of alternative scoring systems, for instance leveraging the difference between the best decoy and best target scores. Following this trend, Emery *et al.* [Emery20, Emery19] proposed the use of multiple decoy database searches and

an antisymmetric score based on the proportion of best decoy matches that were outperformed by the best target matches, similar to other theoretical variations on the knockoff framework [He18a, Gimenez18]. Knockoffs are also inspiring for decoy generation. Essentially, a good knockoff amounts to a random variable following the null hypothesis given its original counterpart. However, this distribution is more difficult to define for mismatching amino acid sequences than for differential expression (where knockoffs easily apply). In fact, this question pervades the problem of decoy fabrication, as a number of methods (reverse, shuffle, De Bruijn, etc.) [Moosa20, Jeong12] have been proposed to allow a trade-off between unrealistic and target-like decoys. However, no consensus has yet emerged. In the knockoff framework, this question boils down to balancing the fit to the exchangeability property, and the assumptions about the null hypothesis. A number of original and inspiring tools have been proposed to achieve this balance, such as the Deep Knockoff approach [Romano19]. This suggests using a generic distribution learner, like a variational autoencoder, with a loss function penalized by the similarity between the target and decoy scores over the training spectra.

Second, and more pragmatically, swapping the default FDR control methods and the use-cases highlights the pros and cons of each approach. We applied BH to peptide identification data [Couté20, Fancello22] and conversely, for differential analysis, we replaced BH by knockoffs section 2.2. Our results concurred and showcased the considerable instability of the competition-based approach relative to null-based approach. This instability is presumably linked to random fluctuations during decoy generation, which directly influence the results of pairwise competitions. In contrast, the overall description provided by the null distribution should be less sensitive to random variations. To cope with these fluctuations, it has been proposed to run multiple TDCs and to average the target and decoy counts to estimate the FDR [Keich19] (an approach that should not be confused with the multiple knockoff approaches mentioned above [Gimenez18]). However, any such averaging strategy comes at an extra computational cost, which is not required when applying the null-based approaches.

This difference in stability also provides a possible explanation for the results presented by Madej *et al.* in [Madej22]. More specifically, the authors proposed to apply the Common Decoy Distribution (CDD) in two distinct ways: BH-CDD (Benjamini-Hochberg CDD) and PP-CDD (PeptideProphet CDD). Technically speaking, both implementations amount to a null-based approach, as both use the CDD as an empirical null distribution. The BH-CDD copes with the lack of available p-values by relying on a simple empirical null model to estimate an overall FDR, whereas PP-CDD, following PeptideProphet approach [Keller02, Choi08], uses a joint modeling of the null and alternative hypotheses to estimate the local FDR distribution [Efron01]. Madej *et al.* reported that, compared to BH-CDD, PP-CDD produces FDR estimations with greater variance. Considering the lack of stability intrinsic to the random generation of fictional variables (whether knockoffs or shuffled sequences), we believe that the minimal data-dependency of BH-CDD (in contrast to PP-CDD, which requires an additional distribution to be estimated from the data) explains its relative stability. As for the reported overconservativeness of BH-CDD, refinements using the PIT (Percentage of incorrect target PSMs [Käll07], a.k.a π_0) are known to be efficient [Keich15, Giai Gianetto16].

Regardless of the strategy applied (PP-CDD or BH-CDD), the distinction between the null- and competition-based approaches sheds an interesting light on the proposal presented in [Madej22]. Indeed, it constitutes a middle-of-the-road approach, as the construction of the CDD involves pairwise competitions, whereas the FDR estimate relies on the empirical null paradigm. Is such an in-between theoretically supported? So far, it does not appear to be, to the best of our knowledge. However, the experimental results reported seem at least partly compliant with a well-calibrated FDR control procedure. This can be explained in a number of ways.

First, when a peptide identification task is conducted with most available database search engines, the null hypothesis is not that of random mismatches, but that of the best mismatches over the entire database (most search engines only return the few best-scoring PSMs, whether

targets or decoys, and classically, only the best one is retained). In other words, even a decoy-only search entails a kind of competition step (in the sense that it amounts to taking the highest of several scores), regardless of its subsequent use (with or without competition against target results). Potentially, after the competition involving the entire decoy database, adding another level of competition with the best target does not significantly alter the distribution. Of course, it should yield a more conservative CDD as only some decoys (those defeating the targets) are considered to describe the null distribution. However, let us recall that depending on how the databases are filtered [Couté20, Fancello22], the TDC-based FDR can be anticonservative. The two errors could thus compensate for each other. An alternative explanation can also be presented: although we currently lack theoretical results, a mixed approach may still hold. It should be remembered that the concept of TDC emerged more than 10 years earlier than the theoretical framework of knockoffs. Similarly, future statistics studies may give grounds for mixing strategies rooted in computational proteomics, where the overall null distribution is determined from a series of pairwise competitions. Although taking a slightly different path, some theoretical attempts to bridge the gap between knockoff- and p-value-based FDRs are already emerging [Etourneau23, He18a, Nguyen20].

In any case, the results reported by Madej *et al.* in [Madej22] are in line with those summarized above: To date in a proteomics application context, competition-based approaches do not appear to us as mature as their null-based counterparts. Owing to the random fluctuations inherent to the generation of fictional features, and the difficulty of ensuring that this generation remains compliant with the mathematical constraints of the underlying theory (which, broadly speaking, is essential to comply with the Equal Chance Assumption), null-based FDR controls should be preferred. More precisely, we tend to promote the BH control, as it is the most stable and the least computationally demanding. Unfortunately, it cannot easily be used with many search engines that do not directly provide p-values. In this context, the most striking application of CDD we envision is a universal method to convert search engine scores to p-values. By definition, p-values distribute like their theoretical quantiles; therefore, the empirical quantiles of the scores attributed by a search engine to a CDD provide a correspondence table between the scores and the p-values for the search engine in question. Subsequent application of the BH procedure then becomes straightforward. In addition to the gains in terms of accuracy and simplicity for FDR control, this would give a common ground to simplify the comparison of the various search engines available in the literature.

2.3.6 Conclusions

In conclusion, the objective underlying the definition of a CDD is as interesting as the new applications it makes possible. However, its practical use raises many questions, mainly because it requires a distinction to be made between two uses of decoy sequences when controlling for the FDR: either to challenge the target sequences in a pairwise competition setting, or to refine the description of the null hypothesis (a.k.a. target mismatches). Doing so in a proteomics context is difficult, since both trends have been tightly intertwined over the past 20 years. The reason for this intermingling is that our community adopted the FDR concept in a progressive and sinuous manner, fueled by a mix of empirical considerations and concomitant theoretical results. Fortunately, the recent advent of knockoff theories is insightful in this respect. In this context, the proposal from Madej *et al.* constitutes a milestone encouraging further investigations in multiple directions. First, casting the CDD principle into a fully empirical null framework (by removing the competition step during CDD construction) would produce a tool that could be extensively used to convert search engine scores into p-values. Second, by providing the means to explore a range of target-decoy strategies, CDD could well become a practical tool to refine our current approaches to FDR control (*e.g.*, stability studies, antisymmetric scores, averaging results from multiple small databases vs using a single large one, π_0 estimate, etc.). Finally, in the longer term, such explorations should contribute to more theoretical investigations; notably attempts to bridge the gap between the statistical rationales of the null- and competition-based approaches, possibly leading to the

emergence of a unified theory.

2.4 Deriving a composite diagnosis score from FDR-controlled biomarker selection

2.4.1 Foreword

This last work, still unpublished, results from a collaboration with David Pérez, another PhD student in molecular biology from EDyP lab, on a project representing the core of his thesis, and for which I am only a contributor. My involvement was motivated by the need to develop a score assessing liver fibrosis severity using a panel of biomarkers either discovered by proteomics experiments or resulting from bibliographical survey. It is an essential follow-up of the proteomics work, as some biomarkers selected using univariate differential analysis and subsequent FDR control may correlate well with one another. Hence, the gain of information by naively combining them may be low.

My involvement in this project was a great opportunity to work directly with experts from other fields and to gain experience, as I ambition to continue my career in interdisciplinary research. This applicative work was yet challenging as the methods to build the prediction model had to fit the constraints of the wet-lab experiments, carried out before my involvement, such as patient variability, size of dataset, potential data leakage *etc.*

This chapter is organized as follows: The next section provides a summary of the article draft to avoid the necessity of reading it. Then, subsection 2.4.3 gathers copied-pasted parts of the article draft or of its supplementary materials, which correspond to my contributions.

2.4.2 Bioclinical context and biomarkers identification

Metabolic dysfunction-associated steatotic liver disease (MASLD) is a disease linked to the spread of obesity, and affects up to 1.9 billion adults, with increasing prevalence [Gadiparthi20, Huang20a]. The most reliable available method to diagnose and characterize MASLD is liver biopsy, which is used to stratify patients with liver steatosis, fibrosis, or inflammation. However, liver biopsy is not without risks, and presents a certain number of limitations in this scenario. Based on its link with mortality [Dulai17], fibrosis is the most urgent variable to be determined in MASLD, and an essential parameter for the medical decision-making process. In particular, the distinction between early stages and advanced stages of fibrosis is critical, as it largely determines whether the patient should receive liver transplantation.

In this context, we are then interested in finding fibrosis biomarkers in blood plasma, to diagnose fibrosis severity (*i.e.*, early vs late stage) in an non-invasive fashion with a blood test. Samples were collected from two cohorts of MASLD patients from different hospitals, one of 160 patients, the other of 200 patients.

A first LC-MS/MS-based discovery proteomics analysis was performed using 158 plasma samples from the Grenoble cohort. The aim was to identify candidate biomarkers with abundance levels making it possible to differentiate early fibrosis from advanced fibrosis. The strategy deployed reliably identified and quantified 235 plasma proteins. The statistical analysis (*limma* test) of the difference between relative abundances of each protein in the samples from patients with early and advanced fibrosis revealed 72 differentially abundant proteins (FDR inferior to 5% according to the Benjamini-Hochberg procedure).

Then, a verification phase enabled to assess whether absolute quantification of target protein (obtained with ELISA assays) differed significantly between samples from the advanced and early

fibrosis groups. To select appropriate candidates for verification, several criteria were applied to the potential biomarkers identified by discovery proteomics: substantial fold-change between the conditions compared, availability and performance of ELISA assays, and predominant expression of the proteins in the liver according to the Protein Atlas database (https://www.proteinatlas.org/). Based on these criteria, from the 72 differentially abundant proteins identified in discovery phase, two proteins were selected for the verification study: ALS and LG3BP. ALS and LG3BP concentrations were determined in the same 158 plasma samples from the Grenoble cohort using ELISA assays. The results showed similar trends in terms of relative abundances for each protein between the discovery and verification studies for the different fibrosis stages.

The validation study aimed to confirm the usefulness of the biomarkers with samples from an independent cohort. To do so, 200 samples from the MASLD patients attending Angers Hospital were analyzed by ELISA. The trends for ALS and LG3BP abundances through fibrosis stages were similar for the two cohorts, whatever the analytical method used. Finally, results obtained for early and advanced stages of fibrosis in the discovery/verification and validation cohorts were statistically compared by applying a Mann-Whitney test. Statistically significant levels were obtained for the discovery, verification, and validation studies for ALS (9.4E-07, 4.4E-07, 3.0E-09) and LG3BP (1.3E-06, 4.1E-07, 2.8E-09, respectively). These levels of significance suggest that the abundances of both proteins can discriminate MASLD patients with early disease from those with advanced liver fibrosis.

2.4.3 Comparison with FibroTest for the non-invasive assessment of liver fibrosis

We then focused on how well ALS and LG3BP differentiated early versus advanced liver fibrosis, examining the biomarkers either independently or in combination. For this analysis, we performed an AUROC calculation on the validation cohort (Angers) using (1) the concentration of ALS measured by ELISA, (2) the concentration of LG3BP measured by ELISA, and (3) the combined concentrations of LG3BP and ALS Figure 2.9. The AUROC values were compared with those obtained with the FibroTest (for the same cohort) as it is widely used and has been tested in several MASLD studies [Lassailly11, Munteanu16, Munteanu18]. FibroTest was initially defined as a panel (*i.e.*, a composite biomarker) of three protein concentrations, one enzymatic activity, one metabolite concentration, age and sex, combined in a logistic regression model and validated on independent cohorts [Poynard05]. In our comparison, ALS and LG3BP provided surprisingly similar performances to the original FibroTest: ALS 0.744 [0.673 – 0.816], LG3BP 0.735 [0.661 – 0.81], FibroTest 0.758 [0.691 – 0.825]. Notably, their 95% confidence intervals (CIs) were found to largely overlap (Figure 2.9). Using the same logistic regression methodology applied to the discovery cohort (Grenoble), we determined the weightings for ALS and LG3BP as a two-protein panel, termed {ALS, LG3BP} on Figure 2.9. With an AUROC of 0.796 [0.731 – 0.862], the {ALS, LG3BP} panel improved fibrosis differentiation as compared to ALS or LG3BP alone, but also compared to the FibroTest panel.

Based on the results obtained, we hypothesized that the combination of ALS and LG3BP with FibroTest (or with at least some of the FibroTest variables) in a multivariate model could improve the overall stratification performance. However, this hypothesis requires careful validation, as modifying a pre-existing panel by adding (or removing) variables must be supported by a precise assessment of each variable's contribution to the overall performance. For this type of validation, nested hypothesis testing has long been relied upon [Krämer86], however, it could not be directly used here because FibroTest was trained on a distinct cohort. To deal with this difficulty, we computed an intermediate mathematical object, called FibroTest RF (re-fitted), which basically consisted in retraining the weightings of the original FibroTest variables using the same original logistic regression, but with data for our discovery cohort (Grenoble). The associated AUROC on the Angers cohort was much higher (0.843 [0.789 – 0.898]) Figure 2.10 than that obtained with the original FibroTest (0.758 [0.691 – 0.825]) (Fig. 4). Despite this apparently dramatic difference, we

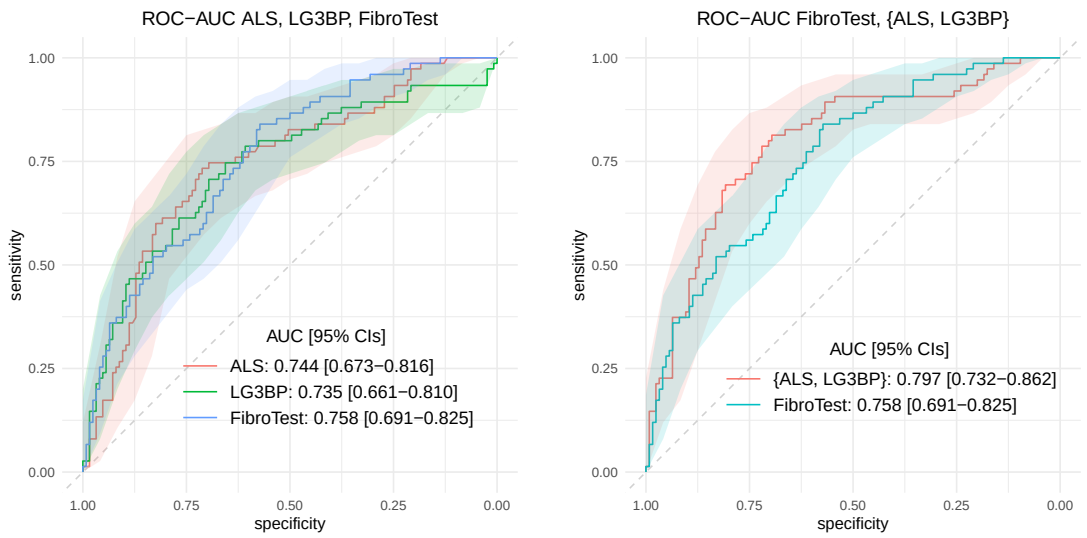

Figure 2.9: Plasma concentration of ALS and LG3BP discriminate early (F0-2) from advanced (F3-4) fibrosis as well as the FibroTest panel; as a 2-protein panel they outperform FibroTest. ROC curves and AUROCs are shown with their respective 95% CIs for ALS/LG3BP ELISA quantifications, original FibroTest score (left). Combined concentrations of ALS and LG3BP compared to original FibroTest (right). Data presented correspond to the Angers cohort. 95% CIs around ROC curves were computed over 2000 stratified bootstrapped replicates of the cohort, using the `pROC` R package, as described in [Carpenter00]. 95% CIs for AUCs were also computed using pROC, but applying the DeLong methodology [DeLong88], as it is an asymptotically exact method; the curves are displayed on two distinct plots for the sake of clarity only, as a result of the extensive overlaps in CIs).

cannot claim that the FibroTest RF represents an improvement upon the original FibroTest as the two were assessed on distinct cohorts. Therefore, FibroTest RF should essentially be interpreted as an abstract mathematical model, providing a baseline for fair comparisons: any modification of the panel yielding a larger AUROC than FibroTest RF can be considered to improve upon the original FibroTest. This conclusion will apply even in the worst-case scenario where the performance difference due to the change of cohort cannot be accounted for. In this context, the results presented in Figure 2.10 support the conclusion that combining previously identified biomarkers with ALS and LG3BP should yield a more powerful panel, better distinguishing between early and advanced stages of fibrosis.

To explain the performance gap between the original and the re-fitted FibroTest, we hypothesized that some of its variables were not true markers of disease severity in the Grenoble and Angers cohorts, and that they could therefore be safely removed from the re-fitted version without loss of performance. We verified this hypothesis by calculating iterative likelihood ratio tests for the Grenoble cohort: at each iteration, we removed the variable the least likely to contribute to the model's performance (see Table 2.2) until only significant variables remained. From this process, we concluded that five variables could be safely removed from the model without decreasing performance, and hence retained only A2M and GGT (gamma-glutamyltransferase). This result is supported by the fact that there is almost no change in AUROC between the re-fitted FibroTest (0.843 [0.789 – 0.898]) and its reduced version A2M, GGT (0.84 [0.785 – 0.895]) (Figure 2.10.)

We then examined whether adding ALS and LG3BP to FibroTest RF or to data related to A2M, GGT improved discrimination between early and advanced fibrosis stages. From the data presented in Figure 2.10, we can see that adding ALS and LG3BP increased the overall AUROCs for these two panels by 0.01 and 0.015, respectively. However, it is difficult to estimate the true gain as CIs considerably overlap. To further assess the comparison, we performed likelihood ratio tests for

these models fitted to the Angers cohort alone (see Table 2.3). Specifically, we tested whether the models including ALS, LG3BP, or both, fit the data significantly better than the models lacking these markers (i.e., FibroTest RF and A2M, GGT models). We obtained p-values of 0.007 and 0.009, respectively, when both ALS and LG3BP were added to these two models. When ALS was added alone, p-values of 0.004 and 0.011 were obtained, respectively, whereas with LG3BP, p-values of 0.066 and 0.0278 were obtained.

By combining ROC curves with nested significance testing, we found that using both previously-discovered biomarkers from the FibroTest alongside ALS and LG3BP produced a more powerful panel to distinguish between early (F0-2) and advanced (F3-4) fibrosis stages. Relying on the Ockham's razor principle, and as we observed no significant effect of the other variables included in FibroTest (*i.e.*, Haptoglobin, Age, Bilirubin, ApoA1 and sex) on our cohorts, we propose a new biomarker panel for liver fibrosis composed of A2M, GGT, ALS, LG3BP. This panel provided an AUROC of 0.855 [0.802 – 0.908] with data for the Angers cohort.

		Model						
		Full FibroTest RF	A2M, GGT, Age, ApoA1, Sex, Bilirubin	A2M, GGT, Age, ApoA1, Sex	A2M, GGT, Age, ApoA1	A2M, GGT, Age	A2M, GGT	Full FibroTest RF
Variables tested for removal	Alpha2m	8.50E-6	8.18E-6	5.79E-6	6.46E-6	2.46E-6	2.67E-7	
	GGT	1.60E-3	1.50E-3	1.42E-3	1.40E-3	2.60E-3	2.70E-3	
	Age	0.271	0.248	0.252	0.226	0.317		0.733
	ApoA1	0.238	0.254	0.249	0.258			
	Sex	0.598	0.6134	0.6435				
	Haptoglobulin	0.734	0.656					
	Bilirubin	0.763						

Table 2.2: P-values from likelihood ratio tests with the lrtest package [Zeileis15] between a given model (in columns), and the same model excluding one variable (given by rows). A low p-value indicates that the variable significantly improves the model's likelihood. We start with the full FibroTest RF model (leftmost column) and iteratively remove the least important variables until only significant ones remained. These tests were performed on the Grenoble cohort. The last column shows the result of tests where all previous variables were removed at the same time.

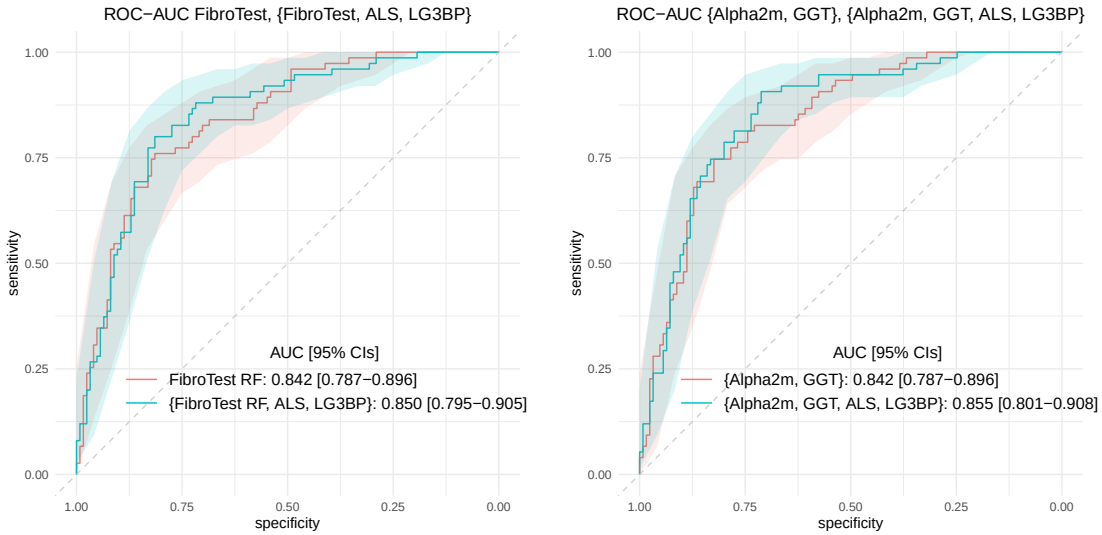

Figure 2.10: Combination of re-fitted FibroTest with ALS and LG3BP improves performance, even with FibroTest reduced to two variables A2M, GGT. ROC curves and AUROCs with their respective 95% CIs are shown for ALS and LG3BP combined with re-fitted FibroTest (FibroTest RF) (left) and A2M, GGT (right), with same methodology as previously. Models were fitted to data from the Grenoble cohort, and ROC-AUROCs were measured using data from the Angers cohort.

		Variable(s) added		
		ALS	LG3BP	ALS & LG3BP
Default model	{A2M, GGT}	0.011	0.028	0.009
	FibroTest RF	0.004	0.066	0.007

Table 2.3: Likelihood ratios tests for A2M, GGT and FibroTest RF improved when ALS, LG3BP, or both were added to the test panel. Improvement was measured using data for the Angers cohort, data are represented as p-values; the lower the value, the more discriminant the test. The null hypothesis for a given p-value in the i-th row and j-th column is that variable(s) in the j-th column do not contribute when added to the i-th model. The alternative hypothesis is that the j-th variable(s) make a significant improvement to the i-th model.

2.5 Closing remarks about FDR and feature selection in proteomics

The three works presented enable us to draw conclusions about the FDR control methods and their limitations in proteomics, as well as about the more general problem of variable selection in high dimensional setting.

Firstly, we highlighted the similarity between knockoff filters and TDC: they have essentially the same scoring method and FDR estimators, and present the same instability with respect to the knockoff filters/decoy database, which seems inherent to competition-based methods. The knockoff theory provides a generic (but unfortunately still approximate) way to generate knockoff with theoretical guarantees, and an indicator assessing the quality of the knockoffs. On the other hand, up to date, no such thing exists for decoy generation, as they are essentially generated with empirical rules. There is thus a need for methods to generate decoys that ensures the Equal Chance Assumption (or more formally, that decoys fulfill the knockoff requirements). Moreover, to allow for an easy use in the proteomics community, an indicator assessing the ECA fulfillment is also necessary.

Secondly, we found out that in knockoff framework, a simple univariate p-value based scoring

was more suited to high-dimensional proteomics dataset than multivariate scores derived from LASSO or others, as proposed in the original knockoff papers. We conclude that the interest of multivariate knockoffs may be reduced when very few samples are available compared to the number of features, such as in proteomics datasets. Yet, classical univariate FDR control methods are limiting to build a composite biomarker, as they do not cope for dependencies between variables. Doing so after the biomarker discovery task requires a variable selection procedure (such as LASSO or ElasticNet [Zou05]) or nested hypothesis tests, which is what we employed on MASLD. Whilst the biomarker combination significantly improved upon results in the MASLD case, it was not guaranteed at all, as the discovery phase was performed in univariate setting (with BH-based FDR control) and did not account for potential correlations with previously known biomarkers. The knockoff approach to build composite biomarkers would have been useful here, but should have been used prior to my involvement in the discovery phase, as to cope for variable dependencies and FDR control concomitantly.

Lastly, the development of a multi-omic imputation method based on the knockoff framework turned out to be an impasse. An original idea was to use to joint original-knockoff covariance matrix generation tool to apply constraints to the covariance matrix of the joint proteomics and transcriptomics datasets, and then use it for imputation (in a similar fashion as knockoff filters are sampled from their conditional distribution with respect to all other knockoffs and original variables). This actually seemed a too difficult path, as the constraints on the covariance matrix used for knockoffs cannot be easily modified, and because of the difficulty to estimate such high dimensional covariance matrices. We can yet draw some lessons about the impact of imputation on differential analysis. In fact, a conservative imputation procedure (in the sense of the subsequent differential analysis) is preferable in proteomics, to avoid to artificially generate false positives. The knockoff framework tells us that generating such variables (that are by definition under the null hypothesis) should be performed without looking at the outcome variable, which is in our case the phenotype or biological/experimental condition of each sample. Thus, we believe it is preferable to impute missing values on the entire dataset at once, instead of separating the different conditions before imputation. This opinion has strongly influenced the strategy underlying Pirat, the imputation algorithm presented in the next chapter. On top of that, it has been recently proven that, in MAR setting, if a variable is missing, its univariate distribution can be also used as a knockoff, thus demonstrating that univariate imputation on a feature is by nature conservative [Koyuncu22] with respect to the differential analysis constraints. The impact of imputation on differential analysis is more deeply addressed in next chapter.

3

A new take on missing value imputation for bottom-up LC-MS/MS proteomics

We present here a novel and original work on missing value imputation for discovery proteomics. This work relies on a statistical model thereafter described, and on biochemical assumptions that allow for quantitative transcriptomic integration.

Sommaire

3.1 Motivations and context of publication

The last two challenges presented in chapter 1 (*i.e.*, MV imputation of LC-MS/MS proteomics data and transcriptomics integration) have motivated the work presented in this chapter.

When starting this thesis, we hoped that transcriptomic integration would enable to overcome to some extents the quantitation limits of LC-MS/MS analyses, which are inherent to the physics of the instruments. The imputation approach presented here, referred to as Pirat, was originally developed towards this goal. However, along its methodological development, the explicit modeling of the instrumental censorship by a missingness mechanism rapidly showed promising results. We therefore dug more deeply towards a "single-omic" imputation strategy, and only proposed a multi-omic extension of it. In the end, the proposed approach to transcriptomic integration is more naive than anticipated, however, we believe it demonstrates the interest of this direction. We thus hope our proof of concept will foster the development of more refined multi-omic imputation approaches, even though the limitations of the central dogma should not be minimized: As discussed in subsection 1.4.2, overall, the sample-wise peptide/transcript linear correlations are weak.

Anyhow, our algorithm, Pirat, demonstrates outstanding performances, either with or without complementary transcriptomic information. The underlying statistical method and its evaluation on a variety of proteomics datasets have been summarized in a preprint (`https://www.biorxiv.org/content/10.1101/2023.11.09.566355v1`) that will be soon submitted to a journal, and that is reproduced in the rest of this chapter.

3.2 Abstract

Label-free bottom-up proteomics using mass spectrometry and liquid chromatography has long established as one of the most popular high-throughput analysis workflows for proteome characterization. However, it produces data hindered by complex and heterogeneous missing values, which imputation has long remained problematic. To cope with this, we introduce Pirat, an algorithm that harnesses this challenge following an unprecedented approach. Notably, it models the instrument limit by estimating a global censoring mechanism from the data available. Moreover, it leverages the correlations between enzymatic cleavage products (*i.e.*, peptides or precursor ions), while offering a natural way to integrate complementary transcriptomic information, when available. Our benchmarking on several datasets covering a variety of experimental designs (number of samples, acquisition mode, missingness patterns, *etc.*) and using a variety of metrics (differential analysis ground truth or imputation errors) shows that Pirat outperforms all pre-existing imputation methods. These results pinpoint the potential of Pirat as an advanced tool for imputation in proteomic data analysis, and more generally underscore the worthiness of improving imputation by explicitly modeling the correlation structures either grounded to the analytical pipeline or to the molecular biology central dogma governing multiple omic approaches.

3.3 Introduction

Bottom-up label-free LC-MS/MS (tandem mass spectrometry coupled with liquid chromatography) stands out as one of the most widely used method to characterize the proteome of biological samples. Analysing the massive amount of MS signals generated by such assays to produce an abundance matrix (a data table storing the abundance of each protein measured in each replicate sample, where samples are listed row-wise and proteins column-wise) requires robust, mathematically well-grounded and fast-evolving software, which development raises computational and theoretical challenges. Notably two difficulties are specific to the bottom-up label-free workflow: First, protein abundances are not directly measured, as sample preparation involves proteolysis (*i.e.*, the cleavage of each protein into several *peptides* using an enzyme like trypsin) and as MS measurements require the peptides to be ionized beforehand (they are then referred to as *precursor ions*, or simply *precursors*). Thus, inferring the identity and quantity of proteins requires two aggregation steps: from precursors to peptides, and from peptides to proteins. Second, raw LC-MS/MS data are hampered by the presence of a large number of missing values (MVs) in the abundance matrices: the overall MVs rate can reach 50% [Lazar16] (whether that is at the level of precursor, peptide, or protein abundances); and at least 50% of peptides usually have at least one MV [Liu21].

The origin of MVs is complex and multi-faceted: since [Karpievitch12], it has been customary to separate them into two types. The first one gathers all the MVs resulting from the various workflow imperfections (such as peptide ionization issues, enzymatic miscleavages, too complex samples, false peptide identifications, *etc.*) and which may affect proteins broadly randomly, regardless of their abundance. The second category corresponds to the censorship mechanism resulting from the instrumental limit: precursors with an abundance below this limit yield MVs. A major issue with this censorship relates to its stochastic and dynamic nature, as the MS range changes along the acquisition process [Vidova17]. As a downside of the dramatic proportion and complex nature of MVs, analysing only fully observed biomolecules (be them precursors, peptides or proteins) is usually not considered, as it would discard too much biologically relevant information.

While it is theoretically safer to conduct the analysis keeping MVs as such [Little19] (*e.g.*, it is possible to identify differentially expressed proteins using models that cope with MVs [Chen14, Ryu14, O'brien18, Goeminne20, Chion23]), many downstream investigation techniques require in practice to impute the MVs first (*i.e.*, to estimate the missing abundances), despite the risk of introducing biases in the original data. This is why, finding methods to accurately impute the manifold of MVs has long been a key challenge of computational proteomic research.

Based on a decade old and vast literature, some consensus about proteomic MV imputation has emerged. First, imputing missing protein abundances has been demonstrated to be suboptimal [Lazar16] as a result of the effect of the peptide-to-protein aggregation operator on missing values. However, whether imputation should be performed at precursor- or peptide-level has not been sorted out yet. Second, it is insightful to distinguish MVs according to the following well-acknowledged statistical categories [Little19]: Missing Completely At Random (MCAR), where the probability that a value is missing does not depend on any observed or missing value in the data; Missing At Random (MAR), where the same probability may only depend on observed values; and Missing Not At Random (MNAR) in any other case. Accordingly, in bottom-up label free proteomics, MVs resulting from the lower instrumental limit are classically assumed to follow an MNAR mechanism [Karpievitch12, Webb-Robertson15, Lazar16], while other MVs, in absence of sufficiently refined MAR-based description, are modelled as MCARs. Third, on the MVs classified as MCARs and on those classified as MNARs, the algorithms providing the best imputations are not the same [Lazar16]. This is notably why their co-existence has motivated the development of meta-imputation tools, *i.e.*, algorithms taking as input one or several imputation methods(s), and delivering as output a more refined imputation result (*e.g.*, [Wei18, Ma20, Giai Gianetto20, Gardner21, Wang22, Chion22]); as well as of diagnosis tools capable of proposing an imputation strategy

tailored to the data specificities [Wang20, Kong22]. Regardless of their practical interest, leveraging those approaches is only possible if classical imputation methods are available in the first place, which explains why we hereafter restrict to them.

Numerous as exhaustive as possible reviews propose comparisons between a host of imputation algorithms (see for example [Lazar16, Karpievitch12, Webb-Robertson15, Jin21, Liu21, Shen22], as well as subsection 3.6.3). Although those works generally concur on the least-accurate approaches, they do not on the most accurate ones, which depend on the dataset or on the experimental design. Moreover, while MCAR/MAR-devoted methods have established their robustness in many scientific domains where MVs have hardly specific behaviors, the MNAR-devoted methods used in proteomics are often simple and univariate, thus poorly informative [Chion23] and leading to larger imputation errors. A notable exception is msImpute [Hediyeh-zadeh23], which very recent publication has unveiled promising results. Contrarily to most anterior methods, msImpute simultaneously tackle MCARs and MNARs, as it estimates the type of each MV, and interpolates MAR and MNAR imputation distributions accordingly. Yet, some parameters (as the barycenter weights or the MNAR assumption) are fixed in a predetermined manner, which may hinder the generalization capabilities of the approach. Pushing the logic a step further, few works [Chen14, Li23] propose to bypass any MCAR/MNAR distinction, as the relevant model should directly estimate the probability of missing depending on the intensity, thus in a full MNAR setting. Among them PEMM [Chen14], has been an important source of inspiration for this work. It proposes a penalized likelihood model involving a random left-censoring mechanism of the mean and covariance matrix of the protein-level data, which estimation yields natural parametric imputations. Unfortunately, this essentially theoretical proposal has shown many practical limitations, which explains its scarce use on real proteomics data.

More broadly, proteomic MV imputation aims at filling gaps in otherwise difficult to interpret biological data, as does multi-omic data integration: This field has elaborated on the natural assumption that a single omic modality contains partial-only information which can be compensated for if multiple omics technologies can be applied to same or related samples. To do so, it is often proposed to estimate a common latent distribution modelling the underlying biological phenomena (classically using either matrix factorization [Argelaguet18, Leppäaho17, Rohart17, Meng19] or deep learning [Zhou20, Barzine20]). Then, the latent model is instrumental for a variety of tasks, such as pathway identification [Rohart17, Argelaguet18, Meng19], sample extrapolation [Rohart17, Barzine20, Zhou20], gene-set analysis [Meng19], and possibly imputation [Flores23], even though none of the available methods was specifically developed (and tested) to fulfil this task in the challenging context of bottom-up label-free LC-MS/MS data. We have however remarked a recent (still unpublished) attempt by Gupta *et al.* [Gupta23] to train graph neural network in a gene-specific manner to impute protein-level MVs from such data, following a logic broadly akin to the one presented here.

Our most important contribution is to propose a concrete route to improve the imputation of LC-MS/MS-based label-free bottom-up proteomic data by leveraging a well-established analytical and biochemical truism: precursors or peptides originating from the same protein should exhibit correlations that are insightful to guess the trend of a protein across several samples. In addition, as insightful correlations can also be exploited between different omics modalities, we have generalized the approach to incorporate quantitative transcriptomic data, hereby opening the path to multi-omic based imputation of proteomic data. Doing so is however not sufficient to tackle the MNAR issue. We have thus implemented these concepts using an imputation algorithm which estimates a single model for both censored and random MVs. Although the model roots on PEMM [Chen14], its estimation is achieved using a completely different strategy, more stable, scalable, and without parameter-tuning or functional approximations, as to fit the concrete constraints of proteomic data analysts. The resulting software, referred to as Pirat (available on GitHub `https://github.com/prostarproteomics/Pirat`), significantly outperforms all other imputation methods on a variety of tasks, datasets, and situations. All our results are reproducible with the code available on GitHub

(https://github.com/TrEE-TIMC/Pirat_experiments).

3.4 Results

3.4.1 Pirat: a novel imputation method for proteomics data

Pirat (standing for *Precursor or Peptide Imputation under Random Truncation*) software works as follows: Firstly, depending on the input data, it creates either Peptide Groups or Precursor Groups. Both are based on protein belonging (details in section 3.6.1) and are abbreviated as PGs, whereas peptides or precursors in a same PG are qualified as *siblings*. For sake of readability, we hereafter refer to peptides only, precursors being mentioned only if they require specific processing. The resulting biochemically informed dependence graph can optionally be enriched with transcriptomic data: for any PG, one then appends the sample-wise abundance vector(s) of transcript(s) (details in section 3.6.1).

Secondly, Pirat estimates over all the PGs a global missingness mechanism:

$$P(M_{i,j} = 1 | X_{i,j}, \Gamma) = \min\left(1, \exp(-\gamma_0 - \gamma_1 X_{i,j})\right) , \tag{3.1}$$

where $\Gamma = \{\gamma_0, \gamma_1\}$ is referred to as the *missingness parameter*, and where $M_{i,j}$ (respectively, $X_{i,j}$) denotes the missingness response (respectively, abundance value) of the i-th sample and the j-th peptide. The missingness parameter is estimated using a kernel regression (details in section 3.6.1) as to fit the missingness pattern of each dataset.

Thirdly, Pirat estimates the same model as PEMM [Chen14] (derived from Rubin's selection model [Little19]), yet using an original approach. Briefly, it aims at maximizing the penalized log-likelihood of observed data and missingness response, which amounts to maximizing the following quantity:

$$\log P(X_{\text{obs}} | \mu, \Sigma) + \log \mathbb{E}_{X_{\text{mis}} | X_{\text{obs}}, \mu, \Gamma} \left[P(M_{\text{mis}} | X_{\text{mis}}, \Gamma) \right] + Q_{K,\lambda}(\Sigma) ,$$

where μ and Σ are the mean and covariance matrix of a multivariate Gaussian distribution, Q is a penalty on Σ and subscripts "obs" and "mis" respectively refer to indices of the missing and observed parts of the dataset (see section 3.6.1 for the detailed mathematical notations). In Pirat, this model is fitted on each PG independently, with respect to μ and Σ, considering Γ is known. Likewise, penalty hyperparameters K and λ are automatically set for each PG in the empirical Bayes framework (details in section 3.6.1). To avoid the approximations of the missingness mechanism (defined in Equation 3.3) that hamper PEMM's results, we propose a tractable and differentiable lower bound of the above quantity. Therefore, the lower bound can be maximized with respect to μ and Σ using a quasi-Newton optimization procedure (L-BFGS [Liu89]). Finally, to decrease computation costs and memory usage, as well as to enforce Σ positive definiteness, Pirat leverages an alternative parameterization of Σ based on its log-Cholesky factorization.

Fourthly, once the model is fitted, we propose to impute the MVs by their conditional mean with respect to observed values and their missingness response. The conditional mean being devoid of closed form, it is computed using standard Monte-Carlo integration [Robert99].

3.4.2 Pirat outperforms state-of-the-art on differential analysis task

We first evaluate our method on several biomarker selection tasks where significantly differentially expressed proteins are sought. We rely on three benchmark datasets for which differentially expressed proteins are known as their relative abundances are controlled. They essentially involve spiked standard proteins in a complex yet constant biological background. These datasets have

been produced with different MS acquisition modes (either data dependent or independent, a.k.a. DDA or DIA), at precursor- or peptide-level, and using different standard mixtures (see Section 3.6.2).

The feature selection is performed using the p-values resulting from the significance testing of the peptides' differential abundance, after imputing with our method (Pirat) as well as with 15 different state-of-the-art or popular methods (described in Section 3.6.3). Although receiving operating characteristic (ROC) curves are classically used to assess feature selection (they can be found in Sup. Mat. Figure 3.5), we have relied on precision-recall (PR) curves (Figure 3.1) instead. They provide broadly similar rankings, but the PR curves display the False Discovery Proportion (FDP), which directly relates to the FDR one often controls in proteomic experiments [Burger18]. More precisely, FDR being classically controlled at 1% or 5%, we present on Figure 3.1 partial PR curves focused on the high precision (low FDP) region to better assess selection performances in this setting. Global area under the entire PR curve (AUCPR) are also given in Table 3.1. On the

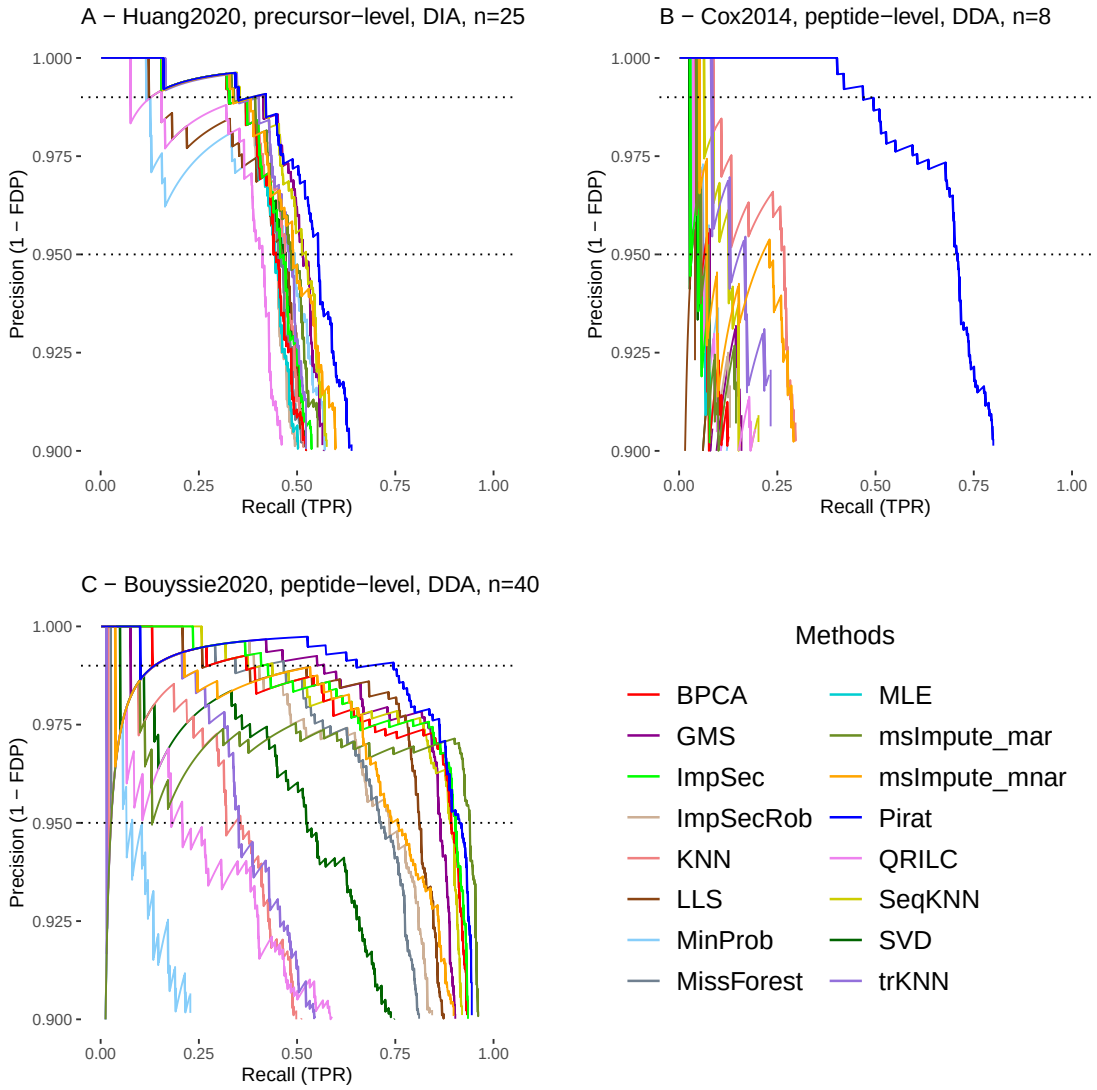

Figure 3.1: Differential abundance PR curves between 90% and 100% precision comparing Pirat to 15 imputation procedures on three benchmark datasets (A - Bouyssié2020, B - Cox2014, C - Huang2020), for which the name of the study, the imputation level (peptide or precursor), the type of acquisition (DIA or DDA), and the total number of replicates (n) are indicated in the subplot titles. The 1% and 5% FDP level are shown by the dotted lines.

	AUCPR % (rank)		
Dataset	A - Huang2020	B - Cox2014	C - Bouyssié2020
BPCA	66.06 (10)	29.96	94.57 (5)
GMS	66.30 (9)	31.25	92.57 (7)
ImpSec	67.20 (7)	31.20	95.08 (2)
ImpSecRob	64.11	29.74	90.47 (9)
KNN	59.33	51.94 (6)	79.94
LLS	62.76	31.48	92.19 (8)
MinProb	72.56 (4)	37.62 (7)	63.96
MissForest	63.81	3.60	88.97
MLE	63.38	34.56 (9)	23.72
msImpute_mar	73.47 (3)	57.37 (5)	94.85 (3)
msImpute_mnar	75.62 (2)	75.68 (2)	94.77 (4)
Pirat	**76.89 (1)**	**87.25 (1)**	**97.22 (1)**
QRILC	68.08 (6)	62.41 (4)	84.66
SeqKNN	66.60 (8)	34.91 (8)	93.81 (6)
SVD	64.54	31.83 (10)	89.66 (10)
trKNN	68.35 (5)	70.49 (3)	81.68

Table 3.1: Global area under the precision-recall curves (and ranking into brackets) comparing Pirat with 15 imputation algorithms on a differential analysis task using three benchmark datasets (Bouyssié2020, Cox2014, Huang2020).

three benchmark datasets, Pirat achieves the highest AUCPR (by a margin of 1.25% to 11.6% with respect to the second-best methods), thus indicating it best preserves the differential abundances without introducing false positives. Precisely, Pirat provides the best PR trade-off on Huang2020 and Bouyssie2020 in the low FDP region, and almost entirely dominates the other methods on Cox2014. Beyond Pirat performances, these experiments are insightful with many respects: Several highly popular methods like MissForest, KNN, SVD, or MLE have poor performances with respect to other less popular yet old methods like TrKNN or ImpSec. Moreover, both msImpute methods (one of the most recent methods, published in April 2023) are very efficient, yet not as much as indicated in their seminal paper [Hediyeh-zadeh23]. The reason is twofold: First, although well ranked in terms of global AUCPR, which concurs with the published results, they are outperformed by many other methods in the low FDP region (for example, SeqKNN exhibits a better PR tradeoff at low FDP than both on Figure 3.1 A and C.) Second, they cannot impute missing values if less than 4 quantitative values are observed, so that the performances drop on datasets with fewer samples, like Cox2014 (more detailed explanations in subsection 3.6.3). Conversely, Pirat clearly outperforms other methods on this dataset, which shows it can safely be applied in context of scarce samples. Most importantly, the lack of stability of other methods with respect to Pirat is worthwhile: Across the 3 datasets, the top scoring methods vary, first because the differences being sometimes marginal, the ranking is sensitive to random fluctuations; but also because of the changes in the MCAR/MNAR proportion. On Bouyssie2020, the MAR/MCAR oriented methods, such as msImpute_mar, ImpSeq, BPCA, and SeqKNN perform best, whereas on Cox2014 and Huang2020 some MNAR oriented methods (like trKNN, msImpute_mnar, MinProb, and QRILC)

display the best performances. Overall, only Pirat remains in the top 3 for all datasets, and it does so with the first rank. Such performances are achievable only because Pirat automatically estimates the missingness mechanism underlying each dataset (see section 3.6.1). In a context where the MCAR/MNAR proportion is often unknown, it is of the utmost importance to have imputation algorithms which are resilient to the nature of missing values, so that a single method can be applied regardless of the data.

3.4.3 Pirat's performances are more stable with respect to MNAR ratio

Although Pirat displays the highest performances in an end-to-end task like biomarker discovery, a detailed analysis of its strengths and weaknesses is of interest to better outline its application scope. To do so, we hereafter refine the evaluations of Pirat by means of mask-and-impute experiments (*i.e.*, missing values are artificially added in real datasets, so that it is possible to measure the difference between the masked values and their imputed counterparts). To provide a baseline to the evaluations, we compare to the best ranked methods according to the previous experiments. However, we were not able to include msImpute_mnar, msImpute_mar, and trKNN despite their relatively good performances, because of their restriction on the number of observed values (see Section 3.6.3). Therefore, the subsequent evaluations compare Pirat with ImpSec (ranked 2 on the MAR dataset Bouyssie2020), SeqKNN (3 times in the top 10), QRILC, MinProb, and BPCA (1 time in the Top 5 and in the Top 10). Like Pirat, they can handle peptides with at least two observed values, while constituting reference methods for the MCAR (ImpSec, SeqKNN, and BPCA) and MNAR (QRILC and MinProb) scenarios.

In this experiment, we rely on datasets from two other LC-MS/MS proteomic analyses, detailed in subsection 3.6.2. The first one (Capizzi2022 [Capizzi22]) is composed of 10 samples processed at the peptide level (16% of MVs). The second one (Vilallongue2022 [Vilallongue22]) contains 8 samples processed at the precursor level (14% of MVs). For both datasets, we only consider the subset of peptides or precursors with no missing value, to which we introduce artificial MVs at a rate equal to that of the original dataset. Finally, we filter out peptides or precursors with one or zero remaining observed values. These artificial MVs are a mix of MCARs and MNARs, where the former ones are generated according to a Probit mechanism [Miao16], as described in Section 3.6.5. Then, we impute the missing values using the algorithms to benchmark and compute their associated Root Mean Square Error and Mean Absolute Error (RMSE and MAE, see Section 3.6.5). We repeat this operation for different MNAR proportions (0%, 25%, 50%, 75%, and 100%) and seeds, leading to the curves of Figure 3.2. We do not display QRILC and MinProb results in Figure 3.2 as their MAE and RMSE is 3 to 4 times higher than that of other methods (full plots available in Sup. Mat. Figure 3.6). For those methods, the difference of performances between mask-and-impute and differential analysis is striking although easily explainable: these MNAR-devoted methods are highly biased towards low abundances and work in a univariate setting (*i.e.*, apart from the knowledge of a ill-defined instrumental limit, no information is used), thus resulting in poor quantitative predictions. Yet, such erroneous predictions may not hinder differential analysis when the low abundance bias is biologically relevant: for example, when a protein is not expressed in a phenotype, imputing its peptides with excessively low values is sometimes harmless.

All the other reference methods tested (PBCA, ImpSeq, and SeqKNN) have similar trends in terms of RMSE and MAE: As MCAR-devoted methods, the associated RMSE and MAE increase with the MNAR proportion (as opposed to QRILC and MinProb, see Sup. Mat. Figure 3.6). Although these three methods are broadly equally performing, ImpSec seems slightly but consistently more accurate, regardless of the MNAR ratio. As for Pirat, even if it is outperformed on Vilallongue2022 in scenarios with a majority of MCAR values, its performances are closer to the top MCAR methods than to MNAR ones on MCAR values. As for the MNARs, it clearly outperforms any other methods. Finally, over the entire range of MNAR/MCAR ratio, Pirat has more stable and more accurate MAE and RMSE trends.

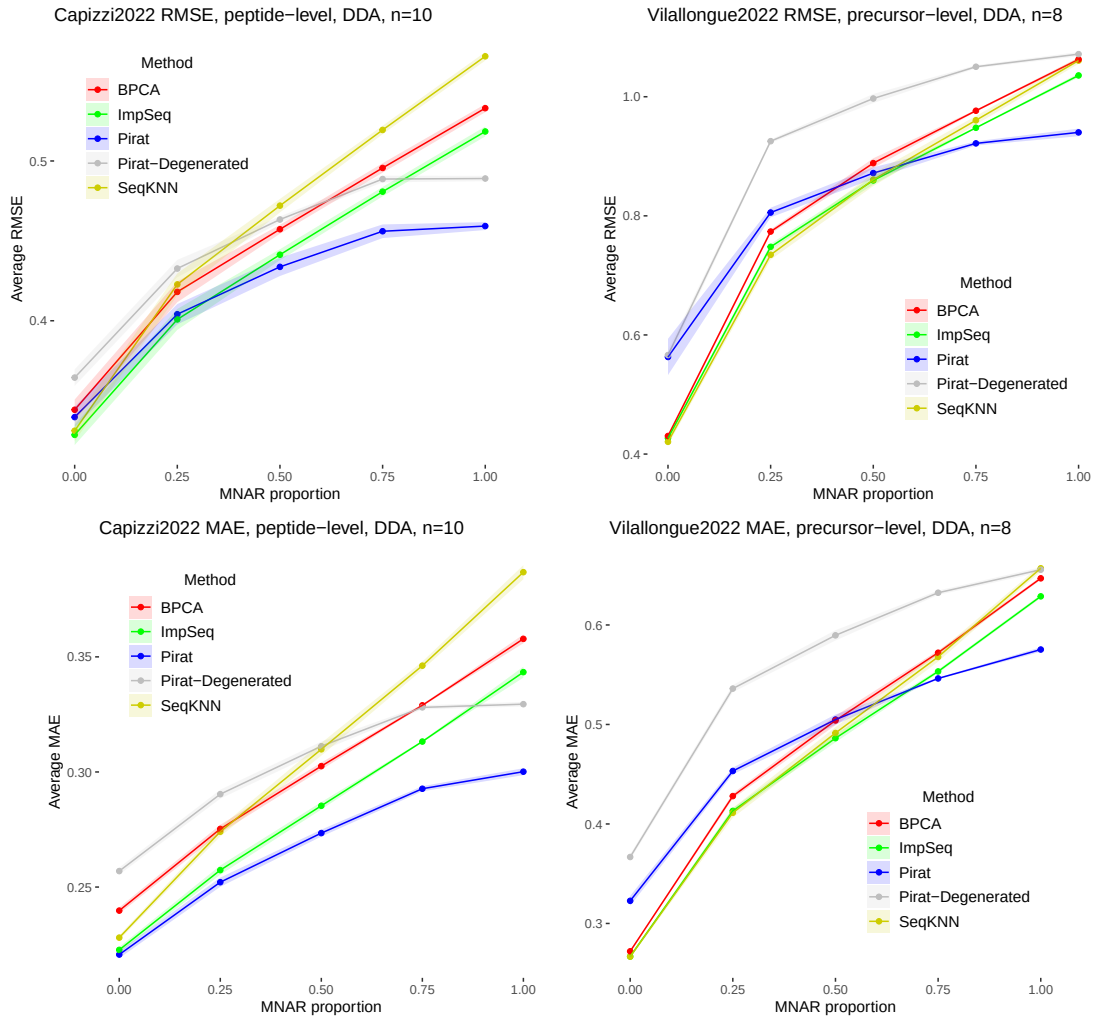

Figure 3.2: Average RMSE (top) and MAE (bottom) of best imputation methods (according to the previous results) in function of the proportion of MNAR values on Capizzi2022 (left) and Vilallongue2022 (right). The imputation level (peptide or precursor), the type of acquisition (DIA or DDA), and the total number of replicates (n) are indicated in the subplot titles. The errors averaged over 5 different seeds and margins correspond to standard deviations.

Although a no-MNAR (or full-MCAR) scenario is not realistic in LC-MS/MS analyses, comparing the associated errors (MAE or RMSE) sheds an interesting light on Pirat's behaviour. When there are few to no MNAR values, the left-censoring model of Pirat loses its interest. In this case, the model boils down to a multivariate abundance one: the estimation can only leverage the correlations between sibling peptides. Conversely, BPCA, SeqKNN, and ImpSeq can leverage the entire feature-wise or sample-wise dependency structure. Hence, the performances on MCARs seem directly related to the magnitude of the within-PG correlations (in the worst case of completely uncorrelated peptides, Pirat would amount to a univariate mean imputation, with likely poor performances). To verify this, we have developed a graphical tool (see Sup. Mat. 3.7.4), which helps assessing the within-PG correlations. Using it on Capizzi2022 and Vilallongue2022 is insightful to understand the difference of performances: With good within-PG correlations, it is not surprising that Pirat performs excellently on Capizzi2022, even with small MNAR ratios. Conversely, with mediocre correlations, Vilallongue2022 is naturally more challenging for Pirat. Pushing the logic further, we evaluated Pirat on a dataset with within-PG correlations hardly larger than between-PG ones (see Sup. Mat. 3.7.4). Without surprise, on this dataset, Pirat is competitive

only in the 100% MNAR scenario. Estimating the precise MNAR ratio of real datasets has always been challenging [Giai Gianetto20, Lazar16]. Even though the amount of MNARs is generally assumed to be significant, we recommend Pirat users to visually assess the within-PG correlations using the tool of Sup. Mat. 3.7.4 as to check whether it will exploit a sufficient amount of common information to correctly impute the few MARs of the dataset. In case both the correlations and the MNAR ratio are anticipated to be low, it might be wiser to prefer a classical algorithm of the state-of-the-art, like for instance ImpSeq. This should notably be the case for peptidomics, top-down, or metabolomics experiments where enzymatic digestion is not used, *de facto* leading to the absence of PGs and thus of within-PG correlations. However, in classical bottom-up proteomics experiments, visualizing no differences between within-PG and random correlations (as in Sup. Mat. 3.7.4) should raise awareness about the quality of the data. Indeed, low within-PG correlations implies that the siblings peptides have possibly different quantitative trends across samples, so that the subsequent protein roll-up will be hazardous, regardless of the imputation quality. Before imputation, it seems important in those cases to question the data acquisition than to opt for an algorithm that does not account for within-PG correlations. Conversely, for datasets where high within-PG correlations are verified using our graphical tool or where classical amounts of MNAR values are anticipated, Pirat will provide unmatched imputation quality.

Finally, Pirat's requirement for solid within-PGs correlations unveils a possible weakness: Weakly-covered proteins (*e.g.*, proteins for which too few peptides are available, which occur in all proteomic experiments) may not be correctly imputed, as no correlation can be leveraged. To assess the performances of Pirat in those situations, we have also considered a degenerated version of Pirat (referred to as Pirat-Degenerated on Figure 3.2), which processes peptides in a univariate way (*i.e.*, as if all PGs were of size 1). Then, Pirat essentially boils down to an imputation by a shifted mean (where the shift depends on the estimated γ_1 parameter, see section 3.6.1). As expected, the imputation performances decrease significantly compared to Pirat: Although the trend is similar to that of Pirat thanks to the censorship model, the broadly constant gap between them confirms the importance of leveraging correlations to boost the imputation performances, while highlighting the risk of lower-quality imputation for weakly-covered proteins.

3.4.4 Improving peptide imputation for proteins with low coverage

We now evaluate various options to best process the peptides of weakly covered proteins. Precisely, we focus on PGs of size 1 (*i.e.*, proteins for which only a single MS evidence is available), thus the most difficult situation for Pirat. We refer to these PGs as *singleton PGs*. As borderline cases, there is no unique satisfactory way to process them: we therefore propose three possible approaches, among which the experimental context (rather than the purely numerical performances) should help choosing. The first one is to adopt a quantitative version of the two-peptide rule often used at peptide identification [Munteanu18], which would lead to filter out singleton PGs with missing values. The second one is to leverage sample-wise correlations for imputation, with a similar logic as other algorithms (*e.g.*, like ImpSeq or BPCA). The third one is to rely on complementary transcriptomic analyses, opening the path to a multi-omic view on imputation.

To evaluate these approaches, we rely on two datasets involving both label-free proteomics and transcriptomics: Ropers2021 [Ropers21] and Habowski2020 [Habowski20]. Ropers *et al.* studied the effect of a synthetic growth switch on *E. coli* by performing paired proteomic and transcriptomic analyses of a wild-type and mutated strain at respectively 2 and 4 timepoints, thus resulting in 6 conditions (3 replicates each). In the mutated strain, protein and mRNA expressions are expected to decrease over time at different rate due to the differences of half-life of mRNA and proteins. As such, the samples have similar compositions leading to a relatively low MV rate at precursor-level (about 15%,) and we expect the transcriptomic-proteomic correlations to be exploitable thanks to the paired design.

In contrast, Habowski2020 results from unpaired transcriptomic and proteomic analyses of

six types of mouse colon cells (3 replicates each for the proteomic experiments and 2 to 5 for the transcriptomic ones). Each condition being a different cell type, the sample composition in proteins varies dramatically between the different conditions. This results in an important amount of precursor-level MVs (which reaches 50%) with values missing on an entire condition (a.k.a. MEC [Wieczorek19]). Note that, compared to Ropers2021, this second dataset is crucial to (1) measure the possible negative impact of uncorrelated transcriptomic data in a multi-omic imputation approach; (2) assess the importance of within-PG correlations to correctly impute poorly observed peptides, containing mostly MNARs (notably MECs). See section 3.6.2 for more details on both datasets.

To evaluate the performance of these three approaches, we rely on a mask-and-impute experiment similar to the previous one, with the additional goal to perturbate the true distribution of MVs the least amount possible. For this reason, we use the entire set of precursors (instead of using precursors with no MVs) and we add only a small amount of MVs (1% of the total number of true MVs in each dataset). This results in a few values that can be used to assess the performance of the approaches. We thus consider only the two extreme scenarios: 0% and 100% MNARs (respectively named MCAR and MNAR scenarios in the following experiments) and repeat this process ten times to assess the overall distribution of errors in each scenario.

Quantitative two-peptide rule

Even though it is technically possible, one should wonder whether it is reasonable to impute missing values for singleton PGs. First, protein abundances for those PGs will directly result from imputation. Consequently, the quality of the imputation will weight a lot in downstream analysis (*e.g.*, differential analysis for biomarker discovery). Second, we cannot exploit the peptide-wise covariances for imputation, and thus suspect deterioration of the imputation for singleton PGs. To confirm this, we compare the median imputation error on singleton PGs with the error of PGs of size greater than two in the MNAR and MCAR settings. Results from this experiment (Figure 3.3, Sup. Mat. Figure 3.9) show that in most cases MVs from singleton PGs are indeed harder to impute than for all other PGs. Akin to the two-peptide rule in protein identification, a simple but effective strategy is to filter out singleton PGs that require missing value imputation to avoid errors in the imputation step to excessively affect downstream analysis. We call this strategy the *quantitative two-peptide rule* or simply the *two-peptide rule*.

Looking more closely at Figure 3.3, the results of Habowski2020 in the MNAR setting are intriguing: we do not observe that the pseudo-MNAR values in this dataset are harder to impute on singleton PGs than on other PGs. A plausible explanation is that the pseudo-MVs of non-singleton peptides (which we compute the error on) in this setting are located on peptides containing mostly MVs. To confirm this, we show in Sup. Mat. Figure 3.10 (A, B, C, D) the histograms of the number of MVs contained in non-singleton-peptides carrying pseudo-MVs. We see that histogram A, representing Habowski2020 MNARs, has a clearly different trend from other scenarios: It is the only situation where pseudo-MVs are, for the vast majority, carried by very scarce peptides containing mostly MECs. Hence, correlations estimated by Pirat are uninformative as they are based on very few values coming from other conditions, which concurs with the drop of performances over non-singleton PGs with respect to singleton PGs in the MNAR setting.

Despite its unquestionable statistical cautiousness, the two-peptide rule approach is sometimes unsatisfactory. Beyond the case of Habowski2020 illustrated above, consider the rather frequent scenario of a confidently identified protein-specific peptide. In such cases, observing distinct observation/missingness patterns hints towards a promising putative biomarker, and consequently, the proposed quantitative version of the two-peptide rule is too stringent. To cope with this, we propose an alternative approach that leverages sample-wise correlations.

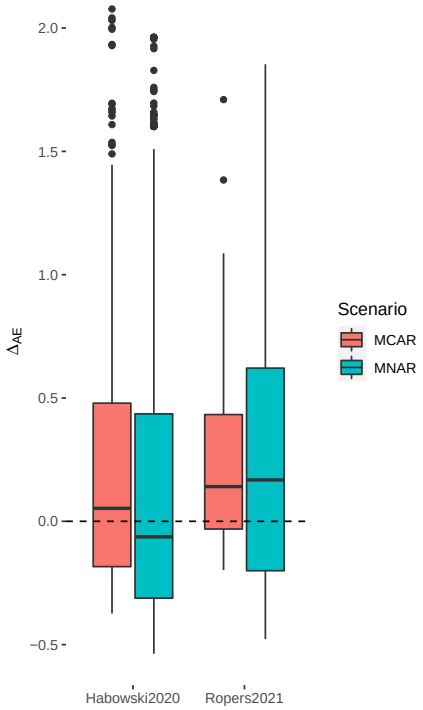

Figure 3.3: Boxplots of the distributions of the differences (denoted as Δ_{AE}) between (i) the absolute errors in singleton PGs and (ii) the median of the absolute errors in all other PGs, after Pirat imputation, for a given dataset (Habowski202 or Ropers2021) and MNAR / MCAR scenario. Hence, for a given dataset and MNAR setting, the part of the boxplot that is above zero corresponds to absolute errors that are greater than the median of absolute errors for non-singleton PGs.

Sample-wise correlations

Pirat is based on the assumption the samples are independently distributed, see section 3.6.1. Yet, other algorithms like ImpSec [Verboven07] assume differently and leverage dependencies between samples. This suggests the following *ad-hoc* strategy: First, impute all non-singleton PGs using Pirat; second, extract sample-wise covariance matrix of the completed data; third, use this matrix to impute the remaining MVs in singleton PGs (more details in section 3.6.1). We refer to this extension as Pirat-S (sample-wise correlation).

Figure 3.4 shows, among others, the absolute errors of Pirat and Pirat-S. Clearly, Pirat-S improves upon Pirat in the MCAR setting for both datasets. This is not surprising: Pirat-S is conceptually close to ImpSeq, which has demonstrated excellent performances on MCARs (see subsection 3.4.3). The improvement is more nuanced in the MNAR setting. Specifically, it seems to deteriorate the performances on Habowski2020, for the same reason as exhibited with the quantitative two-peptide rule, but on singleton peptides: As pseudo-MNAR MVs are mostly carried by empty peptides (see histograms E, F, G and H of Sup. Mat. Figure 3.10), too few values are left to derive informative correlations, notably when they do not originate from the same sample (as with MECs). Although Pirat-S seems to be a safe choice in all other cases, medium-only imputation performances on MNARs with MEC (or similar) patterns reveal a limit of this approach. Indeed, such patterns often result from proteins either absent or too close to the quantitation limit to be consistently measured, so that improving their imputation would *de facto* amount to extending the quantitation range of MS-proteomics. This is why, we propose a third alternative leveraging transcript/protein correlations.

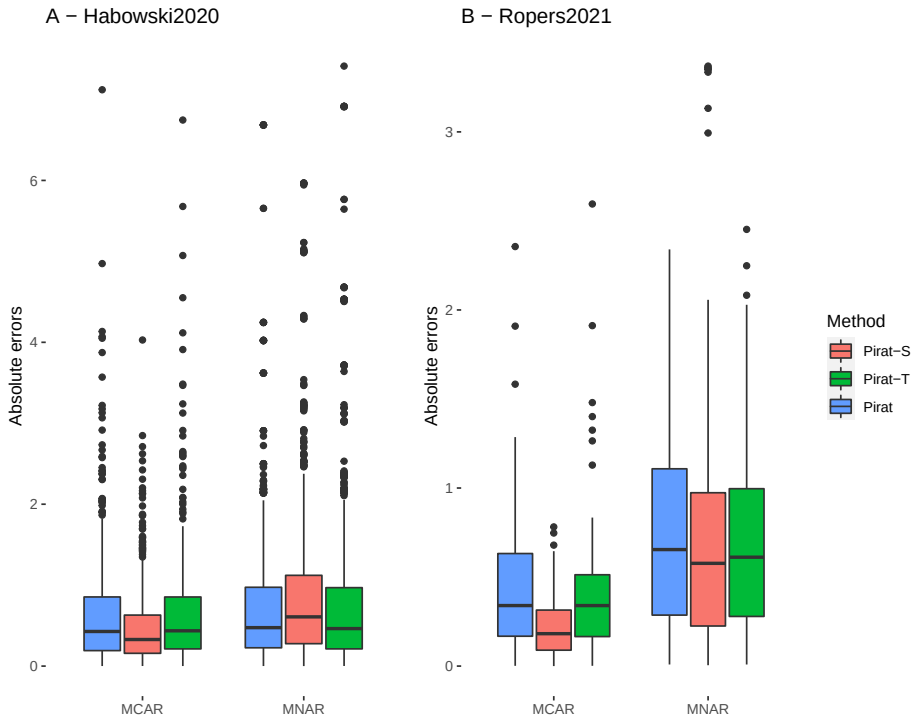

Figure 3.4: Boxplot of absolute errors for singleton PGs in A - Habowski202 and B - Ropers2021 datasets of Pirat, Pirat-S and Pirat-T.

Integrating transcriptomic information

The transcriptome-informed version of Pirat is termed Pirat-T and works as follows: For a given singleton PG, it simply concatenates the transcript abundance vector(s) to the peptide or precursor ones (more details are given in section 3.6.1). Intuitively, the performance increment is expected to depend on the correlations between peptides and transcripts abundances. This is verified on Figure 3.4: the distribution error is shifted downwards on Ropers2021, where proteomic and transcriptomic samples are paired. On the contrary, the differences seem marginal on Habowski2020, where samples are not paired, thus making the transcriptomic patterns hardly exploitable on the proteomic data. Importantly, although transcriptomics is not helpful in this case, it does not deteriorate the performances either. This indicates that even in case a researcher wonders about the transcriptomic data quality or about their correlations with proteomics, it is not riskier to incorporate them for downstream imputation: Pirat-T robustly integrates transcriptomic data for imputation, without over interpreting mRNA variations. When comparing Pirat-S and Pirat-T, it appears the latter do not outperform the former, indicating that in general, sample-wise correlations can be more robustly leveraged than transcriptomic-proteomic ones (see [Fortelny17] for a discussion related to this topic). Although disappointing from a multi-omics integration perspective, this conclusion suffers one noteworthy exception: Habowski2020 on the MNAR setting, and despite the lack of paired design. Following the same logic as before, we conclude that Pirat-T is instrumental for datasets with MECs, as frequently observed when comparing different tissues. As a matter of fact, this confirms the intuition that complementary omic studies can be effective sources of information to explore more safely the proteome beyond the quantitation limit of mass spectrometry.

3.5 Conclusions

In this work, we show that Pirat outperforms previously published imputation methods across various experimental settings. On datasets endowed with differential abundance ground truth labels but with different missingness patterns and replicate numbers, Pirat consistently achieves the best PR trade-off (notably at low false discovery proportion), whereas other competitive methods display less stable and less accurate performances. Likewise, on mask-and-impute experiments, Pirat exhibits lower imputation errors (whatever the metrics, be it RMSE or MAE) for all scenarios with a majority of MNARs. As for the other hardly realistic scenarios with no to few missing values resulting from the instrument censorship, the performances are more variable, yet close to the best ones in the worst cases.

The only weakness our experiments have uncovered relates to Pirat's sensivity to the correlation structure, which has led us to provide the users' community with a diagnosis tool. More specifically, Pirat performances vary with the correlation level amongst sibling peptides. Even though "discordant" peptides among siblings are likely to be observed as a result of unknown post-translational modifications or of poorly quantified peptides [Dermit21], they should be too scarce to affect the correlation distribution. Therefore, too low correlations should raise the experimenter's awareness about the biochemical consistency of the assay. The above case suffers one exception, peptides without siblings (*i.e.*, in singleton PGs), for which the difficulty to impute using Pirat does not relate to the experiment quality. To process them, we propose three alternatives to the original approach: (*i*) the quantitative two-peptide rule, *i.e.*, discarding proteins with a single not fully quantified peptide (*ii*) If transcriptomic data is available, using Pirat-T can be insightful when correlations between peptide and transcript levels are anticipated (notably with paired assays), while it will at least not impair imputation otherwise. (*iii*) If transcriptomic is not available or not relevant for the proteomic study, sample-wise correlations can be leveraged using Pirat-S. However, this requires similar sample compositions to avoid power loss. With this regard, we draw attention to the marginal differences of global performances between Pirat, its two-peptide rule extension, Pirat-T, and Pirat-S: Once averaged on an entire dataset, the imputation error on singleton PGs alone does not weight as much as the experimental design to drive the imputation strategy. We therefore encourage Pirat's users to question their need with respect to proteins with low coverage before opting which strategy to use. Lastly, our experiments have reported Pirat's excellent performances on peptide- and precursor-level datasets, thus ensuring its safe use in both cases. These results unfortunately do not allow us to formulate an educated opinion about which level is most appropriate to impute. Closing this debate would require testing the two approaches with a metric authorizing their comparison.

Considering its performances, we believe future research in proteomics imputation will advantageously leverage Pirat's paradigms. Let us recall them: (*i*) the estimation of a global missingness mechanism inferred from the data; (*ii*) an explicit modeling of the biochemical dependencies that are known to result from the analytical pipeline; (*iii*) the possibility to include other omic measurements that are expected to correlate to the proteomic ones. Pushing further the developments in either of these three trends may require a different mathematical backbone than that of Pirat (*i.e.*, not necessarily involving the penalized likelihood model of Equation 3.4), however, multiple paths are promising: regarding the missingness mechanism, a natural parameter estimator reads as the maximum likelihood estimate (MLE) over all the PGs. However, it requires replacing the optimization procedure described in section 3.6.1 by a joint MLE, which memory and computation cost would be prohibitive. Fortunately, as Pirat processes each PG independently, it is highly parallelizable if acceleration is needed. Also, other missingness patterns like Probit or Logit [Li23] can fit the data, but before investigating them, some guarantees are necessary about their log-likelihood (or any lower bound, see section 3.6.1) being tractable and differentiable. From a biolgical standpoint, databases of known protein-protein interactions (PPI) are a source of supplementary correlation patterns, and incorporating them can be instrumental, notably for singleton PGs (as to find them

pseudo-siblings capable of guiding the imputation). Likewise, known biologically or analytically relevant interactions can also be encoded in the prior of the covariance matrix Σ. For example, the correlations between protein-specific sibling peptides are expected to be larger than those involving either shared peptides, transcripts or PPI, so that it would make sense to adapt accordingly the scaling matrix prior described in section 3.6.1. Finally, Pirat proposes to incorporate transcriptomic data, but other high-throughput technologies provide access to gene-wise or PG-wise quantitative information, which integration into Pirat follows the same logic (broadly speaking, one just appends the quantitative vectors of the various omics into a single matrix). Therefore, Pirat naturally opens the path to more elaborated integrative multi-omic methods to futher tackle the challenge of proteomic data imputation.

3.6 Method

3.6.1 Pirat algorithm

Notations

Before diving into the description of the Pirat algorithm, let us introduce some notations.

- X: a complete matrix of peptide $\log_2$ abundances, with rows referring to samples and columns to peptides, of size $n \times p$.
- $X_{i,j}$: the abundance value in X of i-th row (sample i) and j-th column (peptide j); when i (respectively, j) is replaced by ".", one refers to the to j-th column of X (respectively, the i-th column of X): $X_{.,j}$ thus depicts the peptide vector (respectively, $X_{i,.}$ depicts the sample vector).
- M: an indicator matrix of the same size as X, reflecting whether an abundance value is missing or not. The same indexing notations as of X applies to M ($M_{i,j}$, $M_{.,j}$, $M_{i,.}$).
- $M_{i,j}$: missingness indicator of abundance value of j-th peptide in i-th sample (1 if missing, 0 otherwise).
- $X_{\text{obs}}, M_{\text{obs}}$ (resp. $X_{\text{mis}}, M_{\text{mis}}$): all values of X and M corresponding to observed (resp. missing) abundance values.
- x, m, *etc.*: We use uppercase letters (*e.g.*, $X_{i,j}$ and $M_{i,j}$) when considering random variables and lowercase ones (*e.g.*, $x_{i,j}$ and $m_{i,j}$) when considering their realisations.
- $\{i, \text{obs}\}$ (resp. $\{i, \text{mis}\}$): for a given sample i, set of (i, j) pairs such that $x_{i,j}$ is observed (resp. missing). For clarity, brackets are removed when using this notation as subscript.
- μ: the vector of the theoretical mean abundance of the p peptides.
- Σ: the covariance matrix of the peptide's abundances, of size $p \times p$. The same indexing notation as for matrices X and M ($\Sigma_{j,j'}$ indicates the covariance between peptides j and j').
- Γ: the set of parameters defining the missingness mechanism (see definition in Section 3.6.1).
- I denotes the identity matrix.

Creating peptide groups

We define peptide groups or precursor groups (PGs) according to protein belonging. It is usually assumed that sibling peptides and precursors will quantitatively behave similarly across treatment groups/biological conditions, even though the presence of isoforms, post-translational modifications, or miscleavage can impact these correlations [Dermit21]. Peptides from nested proteins form a unique PG and we duplicate peptides that are shared between PGs. We impute MVs in each PG separately and the multiple imputations obtained for shared peptides are averaged. For sake of readability, peptides or precursors are hereafter simply referred to as peptides, as PGs are processed similarly in both cases.

Model's assumptions

Our model relies on the following four main assumptions regarding the whole peptide dataset (here, X and M refer to all the peptides of all the PGs of the dataset):

1. The $X_{i,.}$ are i.i.d. and normally distributed with parameters μ and Σ.

2. The missingness mechanism is self-masked for all peptides [Sportisse20], (*i.e.*, the probability of an abundance value being missing, knowing the entire sample's abundances, only depends on the abundance itself):

$$P(M_{i,j} = 1|X_{i,.}) = P(M_{i,j} = 1|X_{i,j}) \quad \forall i, j \tag{3.2}$$

3. The missingness responses for each peptide of a sample are independent conditionally to their abundance [Sportisse20]:

4. The missingness mechanism can be written as [Chen14] :

$$P(M_{i,j} = 1|x_{i,j}, \Gamma) = \min(1, \exp(-\gamma_0 - \gamma_1 x_{i,j})), \text{ where } \Gamma = \{\gamma_0, \gamma_1\} \text{ with } \gamma_1 \geq 0; \tag{3.3}$$

   We assume here that there is a minimum detection threshold $-\gamma_0/\gamma_1$ below which no peptide can be quantified, and that the probability for a peptide to be missing exponentially decays with log-intensity.

These assumptions ensure identifiability of μ, Σ, and Γ parameters in univariate and multivariate cases, as demonstrated in [Sportisse20, Miao16].

Estimation of the missingness parameters

The set of parameters Γ over the whole peptide dataset is estimated as follows:

1. Sort the peptides by their observable mean, and denote by α_i the mean of the i-th peptide i in the ordered vector α;

2. For $i = \lceil \frac{k+1}{2} \rceil, \ldots, p + 1 - \lfloor \frac{k+1}{2} \rfloor$ (where $\lceil . \rceil$ and $\lfloor . \rfloor$ denotes the upper and lower rounding), compute the rolling average of the missingness percentage y_i over the ordered peptides, with a window of size k; which leads to $p - k + 1$ points (α_i, y_i);

3. Fit a linear model on the points (α_i, y_i) by ordinary least squares, and set the parameters of the missingness mechanisms 3.3 as the coefficients obtained.

At $k = 10$, we observe a clear and smooth decreasing trend $\log(y)$ with respect to α on two real datasets (Habowski2020 and Ropers2021) and a satisfying R^2 (representing goodness of fit, see Sup. Mat. Figure 3.11). This value is thus used in all experiments. The parameters γ_0 and γ_1 are hereafter considered fixed and known.

Penalized likelihood model

Considering the missingness parameters Γ known, we now propose to maximize the joint log-likelihood $\mathscr{L}$ of the observed values and of the missingness response (following Rubin's selection model [Little19]) with respect to the Gaussian parameters μ and Σ, iteratively for each PG:

$$(\hat{\mu},\hat{\Sigma}) = \underset{\mu,\Sigma}{\operatorname{argmax}} \quad \mathscr{L}(\mu,\Sigma|X_{\text{obs}},M,\Gamma) + Q_{K,\lambda}(\Sigma), \tag{3.4}$$

where X and M hereafter refer to the abundance matrix and missingness response of a single PG, and where $Q_{K,\lambda}(\Sigma)$ is a penalty term on Σ with hyperparameters K and λ (see Section 3.6.1), as proposed in [Chen14]. Hence, maximizing this penalized log-likelihood amounts to maximizing the posterior distribution of parameters μ, Σ, and Γ. Considering previous assumptions, the log-likelihood of the selection model decomposes as:

$$\begin{aligned}\mathscr{L}(\mu,\Sigma|X_{\text{obs}},M,\Gamma) &= \log P(X_{\text{obs}}|\mu,\Sigma) + \log P(M_{\text{obs}}|X_{\text{obs}},\Gamma) + \log \mathbb{E}_{X_{\text{mis}}|X_{\text{obs}},\mu,\Gamma}\left[P(M_{\text{mis}}|X_{\text{mis}},\Gamma)\right] \\ &= \log P(X_{\text{obs}}|\mu,\Sigma) + \log P(M_{\text{obs}}|X_{\text{obs}},\Gamma) \\ &\quad + \sum_{i}^{n} \log \mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma}\left[\prod_{j\in\{i,\text{mis}\}} P(M_{i,j}=1|X_{i,j},\Gamma)\right].\end{aligned} \tag{3.5}$$

Penalty over Σ

We estimate Σ and μ for each PG independently by optimizing Equation 3.4. Therefore, the penalty hyperparameters λ and K must be tuned automatically (as many times as PGs). To do so, we rely on a Bayesian interpretation of the penalty term $Q_{K,\lambda}(\Sigma)$: it can be viewed as an inverse Wishart prior of the covariance matrix Σ. This penalty can be rewritten as:

$$\begin{aligned}Q_{K,\lambda}(\Sigma) &= \frac{1}{2}2K\log(|\Sigma|) + \frac{1}{2}tr(2\lambda I\Sigma^{-1}) \\ &= \log[p_{\mathscr{W}^{-1}(2K-p-1,2\lambda I)}(\Sigma)] + C,\end{aligned} \tag{3.6}$$

where $p_{\mathscr{W}^{-1}(2K-p-1,2\lambda I)}$ denotes the density of an inverse Wishart distribution with parameters $(2K-p-1, 2\lambda I)$, and where C is constant with respect to Σ. Hence, K is related to the number of degrees of freedom in Σ estimate, and λ is related to the scaling matrix $2\lambda I$.

This leads us to empirically estimate the hyperparameters of each PG from the set of all fully observed peptides by relying on the empirical Bayes framework [Efron72] where we leverage that, in the univariate case, the inverse Wishart distribution boils down the inverse-gamma one. Specifically, we compute their empirical variance and estimate the parameters of an inverse-gamma distribution (denoted α and β) through an MLE. Finally, by setting $K = \alpha + p$ and $\lambda = \beta$, the marginal prior distribution of each diagonal element of Σ (the variance of each peptide) is an inverse-gamma with parameters (α,β). Hence, the peptide's variances have the same prior distribution among all PGs (this property is instrumental as before looking at any data, one would not expect the variance of a peptide to depend on the PG it belongs to). Without relaxing this property, we increase α by a factor 2 to better constrain the estimation of the covariance matrix (following [Chen14], which showed in various scenarios that increasing K improves the estimation of Σ).

Deriving a lower-bound of the penalized log-likelihood to estimate μ and Σ

A main issue regarding the estimation of μ and Σ by maximizing Equation 3.5, is that the expectation from the log-likelihood is not analytically tractable for any missingness mechanism. To compute it, Chen *et al* [Chen14] approximated Equation 3.3 as:

$$P(m_{i,j}=1|x_{i,j},\Gamma) = \exp(-\gamma_0 - \gamma_1 x_{i,j}) \tag{3.7}$$

and set any observed value below $-\gamma_0/\gamma_1$ as missing. We claim that this approximation makes the log-likelihood unbounded, and hinders convergence in some cases, according to the following results. Indeed, using this approximation, we have (see [Chen14], Supp. Mat. Appendix D) $\forall i \in \{1,...,n\}$:

$$\mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma}[P(M_{i,\text{mis}}|X_{i,\text{mis}},\Gamma)] = \exp(-\gamma_0|\{i,\text{mis}\}| - \gamma_1 \mathbf{1}^T\tilde{\mu}_i + 1/2\cdot\gamma_1^2\mathbf{1}^T\tilde{\Sigma}_i\mathbf{1}), \tag{3.8}$$

where $\tilde{\mu}_i, \tilde{\Sigma}_i$ are the parameters of the normal distribution $X_{i,\text{mis}}|X_{i,\text{obs}}$. The above quantity diverges towards $+\infty$ if any coefficient of $\tilde{\Sigma}_i$ tends towards $+\infty$, which is not consistent with the definition of a probability distribution.

However, using Jensen's inequality, and the original missingness mechanism from Equation 3.3, we can derive a tractable and differentiable lower bound of the likelihood that can be optimized. Concretely, to estimate μ and Σ for a given PG, we propose to maximize the following lower bound, $\forall i \in \{1,...,n\}$:

$$\begin{aligned}\log \mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma}[P(M_{i,\text{mis}}|X_{i,\text{mis}},\Gamma)] &\geq \mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma}\Big[\sum_{j\in\{i,\text{mis}\}}\log P(m_{i,j}=1|x_{i,j},\Gamma)\Big] \\ &\geq \sum_{j\in\{i,\text{mis}\}}\mathbb{E}_{X_{i,j}|X_{i,\text{obs}},\mu,\Sigma}\big[\log P(m_{i,j}=1|x_{i,j},\Gamma)\big]\,. \end{aligned} \tag{3.9}$$

Let $W_i \sim X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma \sim \mathcal{N}(\tilde{\mu}_i,\tilde{\Sigma}_i)$, let w_i be the realisation of W_i, and let n_i^{mis} be the number of missing values in the i-th sample. Using the parametric model from Equation 3.3, $\forall j \in \{i,\text{mis}\}$:

$$\begin{aligned}\mathbb{E}_{X_{i,j}|X_{i,\text{obs}},\mu,\Sigma}\big[\log P(m_{i,j}=1|x_{i,j},\Gamma)\big] &= \int_{-\frac{\gamma_0}{\gamma_1}}^{+\infty}(-\gamma_0-\gamma_1 w_{i,j})\phi(w_{i,j}|\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j})\,\mathrm{d}w_{i,j} \\ &= -\gamma_0\left(1-\Phi\left(-\frac{\gamma_0}{\gamma_1}\Big|\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}\right)\right) \\ &\quad -\gamma_1\int_{-\frac{\gamma_0}{\gamma_1}}^{+\infty} w_{i,j}\phi(w_{i,j}|\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j})\,\mathrm{d}w_{i,j}\,,\end{aligned} \tag{3.10}$$

where $\phi(\cdot|\mu,\sigma)$ (resp. $\Phi(\cdot|\mu,\sigma)$) denotes the (resp. cumulative) distribution function of a normal distribution with parameters μ,σ. Then, using properties of truncated normal distribution, we have:

$$\begin{aligned}\mathbb{E}_{X_{i,j}|X_{i,\text{obs}},\mu,\Sigma}\big[\log P(m_{i,j}=1|x_{i,j},\Gamma)\big] &= -\left(1-\Phi\left(-\frac{\gamma_0}{\gamma_1}\Big|\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}\right)\right) \\ &\quad\times\left(\gamma_0+\gamma_1\mathbb{E}\left[W_{i,j}\Big|W_{i,j}\geq-\frac{\gamma_0}{\gamma_1}\right]\right) \\ &= -\left(1-\Phi\left(-\frac{\gamma_0}{\gamma_1}\Big|\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}\right)\right)(\gamma_0+\gamma_1\tilde{\mu}_{i,j}) \\ &\quad -\gamma_1\tilde{\sigma}_{i,j}^2\phi\left(-\frac{\gamma_0}{\gamma_1}\Big|\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}\right)\,.\end{aligned} \tag{3.11}$$

The expression of Equation 3.12 is differentiable with respect to parameters $\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}$. These two are the parameters of the conditional distribution of $X_{i,j}$ with respect to $X_{i,\text{obs}}$. They are obtained by the linear combination of μ and the Schur complement of the covariance matrix of observed values in i-th sample. Schur complement is differentiable with respect to Σ so that $\tilde{\mu}_{i,j}$ and $\tilde{\sigma}_{i,j}$ are differentiable with respect to Σ and μ. Hence, we can use any automatic differentiation tool, combined with optimization algorithm, to maximize the following lower bound of Equation 3.4:

$$\begin{aligned}\mathscr{L}(\mu,\Sigma,|X_{\text{obs}},M,\Gamma) &= \log P(X_{\text{obs}}|\mu,\Sigma)+\log P(M_{\text{obs}}|X_{\text{obs}},\Gamma)+ \\ &\sum_i^n\sum_{j\in\{i,\text{mis}\}}\mathbb{E}_{X_{i,j}|X_{i,\text{obs}},\mu,\Sigma}\big[\log P(m_{i,j}=1|X_{i,j},\Gamma)\big]+Q_{K,\lambda}\,.\end{aligned} \tag{3.12}$$

Specifically, we use L-BFGS (a quasi-newton method) combined with Armijo backtracking line-search, implemented with Pytorch [Shi21].

To ensure Σ positive-definiteness during the optimization process, we re-parametrize Σ by its log-Cholesky factorization [Pinheiro96]. Finally, to avoid ill-conditioned matrix that could cause numerical instabilities, we apply Tikhonov regularization method and subsequently add a fixed value $\varepsilon_\Sigma = 10^{-4}$ to the diagonal of Σ. Doing so does not impact on the optimum solution, as the minimum variance observed in all the datasets processed so far has been $\geq 10^{-3}$.

Imputation

Once all parameters μ, Σ, and Γ are estimated, we impute missing values by their conditional mean. To do so, we use the following result [Chen14]:

$$\mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},M_{i,.},\mu,\Sigma,\Gamma}[X_{i,\text{mis}}] = \frac{\mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma}[P(M_{i,\text{mis}}|X_{i,\text{mis}},\Gamma)X_{i,\text{mis}}]}{\mathbb{E}_{X_{i,\text{mis}}|X_{i,\text{obs}},\mu,\Sigma}[P(M_{i,\text{mis}}|X_{i,\text{mis}},\Gamma)]} \tag{3.13}$$

The distribution on which the left expectation is computed is not accessible, however, the distributions of the right expectations are Gaussian. Hence, we use Monte-Carlo integration to estimate them and compute the conditional mean of $X_{i,\text{mis}}$. Note that, using the same approach, the order 2 moment of $X_{i,\text{mis}}$ can be computed by simply squaring it in the above equation, which gives access the variance of $X_{i,\text{mis}}$.

Sample wise correlation to impute PGs singleton

We propose in Pirat-S to leverage sample-wise correlations to impute peptides of singleton PGs. To do so, we first estimate missingness parameters Γ and hyperparameters K and λ on the whole dataset and impute only PGs of size > 1 with the classical version of Pirat, described above. Second, we compute the empirical sample mean and sample-wise covariance matrix of the transposed imputed part of the dataset. Finally, we impute the remaining MVs by their conditional mean with respect to observed values, only using the observed values and the empirical sample mean and covariance matrix.

Transcriptomic integration

Pirat-T extends Pirat by enabling the integration of transcriptomic quantitative information, when available, as to guide peptide imputation. To do so, it requires a dataset of $\log_2$ mRNA expressions of samples from the same phenotype(s) as the proteomic ones. However, increments in imputation performances may require paired transcriptomic and proteomic samples. It also requires a correspondence table between transcripts and PGs, for instance, based on their original gene(s). In this work, transcriptomics is essentially used to improve singleton PG imputation (see section 3.4.4). However, Pirat's methodology is more versatile and the packaged code makes it possible to tune which PGs are imputed using complementary transcriptomic information, depending on their size.

Concretely, Pirat-T works as follows: for each PG, it integrates all the transcripts (*i*) for which non-zero values are measured in at least two conditions, and (*ii*) which are associated to at least one of the proteins of the PG. In the case proteomic and transcriptomic samples are paired, the log-count vectors of the transcripts are simply appended to the PGs. Otherwise, a condition-wise mean log-count vector is used instead. *In fine*, each mRNA log-count vector is processed like an additional (fully observed) peptide, and thus contributes to imputation depending on its correlation with the sibling peptide(s).

3.6.2 Datasets

To conduct our experiments, we rely on the following publicly available datasets:

- **Cox2014** [Cox14]: a mixture of 48 recombinant human proteins that are available as an equimolar mixture (UPS1) or mixed at defined ratios spanning 6 orders of magnitude (UPS2) with respect to UPS1 (10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}). UPS1 and UPS2 are separately digested and resulting peptides are spiked into a *E. coli* lysate, resulting in two distinct experimental conditions. Each condition is analysed four times in data dependent acquisition (DDA, *i.e.*, a single precursor is iteratively selected for fragmentation and identification by the mass spectrometer according to its response level in the first MS acquisition), resulting in four technical replicates per conditions. We use the file `peptide.txt` available at repository PXD000279, containing peptide-level abundances.

- **Bouyssie2020** [Bouyssié20]: a mixture of UPS1 proteins spiked at different concentrations in the same yeast lysate. The dataset contains 10 conditions with 4 technical replicates each (each condition corresponding to a different UPS1 quantity, namely, 0.01-0.05-0.1-0.250-0.5-1-5-10-25-50 fmol, for 1μg of yeast lysate). Hence, unlike Cox2014, within each condition, all the UPS proteins are spiked in with identical concentrations. Replicates are analysed in DDA mode. We use peptide intensities from `allsamples_sum.xlsx` file available at repository PXD009815.

- **Huang2020** [Huang20b]: the UPS2 mixture spiked in tissue lysates from 25 mouse cerebellum samples. Five conditions (with 5 technical replicates each) are generated from spiked UPS2 proteins in known and different concentrations (namely 0.75-0.83-1.07-2.04-7.54 amol/μl) into mouse cerebellum samples. The LC-MS/MS runs are acquired by the data independent method (DIA, *i.e*, a set of precursors in a given mass to charge range are iteratively selected for fragmentation and identification, in order to cover the whole MS acquisition, regardless of the measured intensities). We use the published Spectronaut results in `Spike-in-biol-var-OT-SN-Report.txt` available at repository PXD016647, and imputed MS1 quantification at precursor level.

- **Capizzi2022** [Capizzi22]: a study on Huntington's disease effect on axonal growth in mouse. Two conditions with 5 biological replicates each are compared using LC-MS/MS in a data dependent acquisition mode (DDA), representing wild-type vs genetically modified model. Details of sample preparation can be found in the original article. The accession number for this dataset is PXD023885. Raw data is processed in the same manner as described in the original paper to obtain peptide-level quantification table.

- **Vilallongue2022** [Vilallongue22]: a study on the influence of injury on visual targets in mouse. Five different tissue types are analysed by LC-MS/MS in DDA acquisition mode in independent quantification tables, with each time a control and an injured condition (4 biological replicates each). To test our method, we choose the two tissues that show the best and worse within-PG peptide correlations among the five (see 3.7.4): respectively, the suprachiasmatic nucleus tissue (SCN), and the superior colliculus (SC). The details of sample preparation can be found in the original article. The accession number for this dataset is PXD029325. Raw data is processed in the same manner as described in original paper to obtain precursor-level quantification table.

- **Habowski2020** [Habowski20]: a study on the differentiation of mouse colon epithelial cells. Stem cells and 5 differentiated cells are analysed at proteomic and transcriptomic levels in an unpaired manner. Three biological replicates are used for each cell type in proteomic analyses. The transcriptome dataset is produced by Illumina paired-end stranded sequencing. The proteome dataset is produced by LC-MS/MS in DDA acquisition mode.

  - **Transcriptomic data processing**

We download the raw sequencing data from GEO (GSE143915). Preprocessing and quality control are performed using the Trimmomatic [Bolger14] and FastQC tools respectively. Trimmed reads are aligned to the mouse GRCm39 genome assembly (filename given in Sup. Mat. subsection 3.7.2) by the STAR mapping software (version 2.7.8a) [Dobin13] provided with the Ensembl gene model given in Sup. Mat. subsection 3.7.2 for mapping splices. Read counts are generated using HTSeq (version 0.9.1; option -s no) [Anders15]. DESeq2 version 1.22.1 [Love14] is used to generate normalized counts.

- **Proteomic data processing**

  We download the raw spectra from ProteomeXchange (accession ID: PXD019351). Peptides and proteins are identified using Mascot (version 2.7.0.1, Matrix Science) searching the Ensembl protein database for *Mus musculus* GRCm39 (filename given in Sup. Mat. subsection 3.7.2) appended with an in-house classical contaminant database. Mascot search is performed with the following parameters: trypsin/P as enzyme and two missed cleavages allowed; precursor and fragment mass error tolerances at 10 ppm and 20 ppm, respectively. The following peptide modifications are allowed during the search: Acetyl (Protein N-term, variable) and Oxidation (M, variable). The Proline software [Bouyssié20] is used to validate identifications: conservation of rank 1 peptides, peptide length ≥ 7 amino acids, false discovery rate (FDR) of peptide-spectrum-match identifications $< 1\%$ as calculated by Benjamini-Hochberg procedure and minimum of 1 specific peptide per identified protein group. Peptide ion intensities are calculated from extracted mass spectrum intensities of all peptides and normalized using variance stabilizing transformation with Prostar [Wieczorek17]. A more detailed description of the parameters used in the quantification step are provided in Supplementary Table 1.

- **Ropers2021** [Ropers21]: a study on growth-arrested and wild type *E. coli* cells carrying a plasmid for glycerol production at various time points. A wild type and a modified strain are analysed in a paired manner at proteomic and transcriptomic levels. The modified strain is analysed at four different time points, resulting in four different conditions, and two others from the wild-type on the two first time points. Three biological replicates are made for each condition. The transcriptome dataset is produced by stranded sequencing on the Ion S5 using the Ion 540 chip. The proteome dataset is produced using LS-MS/MS in DDA acquisition model.

  - **Transcriptomic data processing**: Same procedure than for Habowski2020 is applied. GEO accession number is GSE168336. Trimmed reads are aligned to the reference *E. coli* K12 substrain MG1655 genome (Genbank assembly accession: GCA_000005845) provided with the Ensembl gene model given in Sup. Mat. subsection 3.7.2 for mapping splices.

  - **Proteomic data processing**:

    The same procedure than for Habowski2020 is applied, except for the following: Raw spectra are downloaded from ProteomeXchange (accession ID: PXD024231). Peptides and proteins are identified by searching the same Ensembl protein database used in transcriptomic analysis (for this Ensembl genome assembly of *E. coli* the protein identifier is identical to the transcript identifier), appended with an in-house classical contaminant database, the plasmid pCL1920 protein sequences and the protein sequences of the GPD1 and GPP2 yeast genes which were cloned into it (Uniprot accessions: Q00055 and P40106). Carbamidomethyl (C, fixed) peptide modification is additionaly allowed during search. Peptides with length ≥ 6 amino acids are conserved

at validation of identifications. A more detailed description of the parameters used in the quantification step are provided in Supplementary Table 2.

On all datasets, we remove peptides or precursors having strictly less than 2 observations. Finally, we apply $\log_2$ transformation both on peptide or precursor intensities and transcript normalized counts (when available).

3.6.3 List of competitive methods

To benchmark Pirat, we use 15 state-of-the-art imputation methods, many of which have already been pinpointed by NAguideR evaluation software [Wang20]: Bayesian Principal Component Analysis (BPCA) [Oba03] and SVD [Troyanskaya01] parametrized with a number of components equal to the number of compared conditions in each dataset; GMS [Li20] with 3 fold cross-validation for the tuning of its inner parameter (termed λ); K-nearest neighbors (KNN) [Troyanskaya01] SeqKNN [Kim04], trKNN [Shah17] with the number of neighbors k set to 10; ImpSeq [Verboven07] and its outlier oriented version ImpSeqRob [Branden09], Local Least Square (LLS) [Kim05], MinProb [Lazar16], MissForest [Stekhoven12], MLE [Love14], msImpute in MAR and MNAR versions [Hediyeh-zadeh23], and QRILC [Lazar16] with default parameters. MissForest and LLS were tested with an input format where peptides are in columns and samples in rows, contrarily to all other methods. Amongst those methods, the following ones are specifically devoted to low abundance censored values: MinProb, msImpute_mnar, QRILC, TrKNN.

While most methods require only 2 observed values (sometimes less) per peptide, GMS and trKNN require at least 3 of them, and both msImpute versions, 4 of them. This is an issue for two reasons: First, because, on controlled datasets, the biological variability is so small that not imputing those peptides before testing for differential abundance does not hamper the results, hereby artificially boosting the feature selection performances with respect to real-life cases. Second, because observing only two values is common for peptides nearby the detection limit, so that a concrete assessment on the MNAR imputation error is not possible with these algorithms. This is why, regardless of their performances on the biomarker selection task (see Section 3.4.2), GMS, trKNN, msImpute_MNAR, and msImpute_MAR could not be considered in the mask-and-impute experiments.

Likewise, three remaining state-of-the-art methods could not be included in the benchmark for various reasons: (*i*) Penalized Expectation Maximization for Missing Values (PEMM) [Chen14], which had originally been validated on a simulated dataset with 100 proteins and could not scale up to an entire real-life peptide/precursor dataset as a result of its prohibitive computational cost; (*ii*) The PIMMS methods [Webel23] which are based on deep-learning approaches, and which consequently require large cohorts (more than 50 samples) to be fully efficient; (*iii*) ProJect [Kong23], a general purpose omics imputation method which code has not yet been commented, structured and packaged at the time of our evaluations, so that we could not make it work on the benchmark data. Finally, few elder methods with previously demonstrated poor performances (*e.g.*, zero imputation, mean imputation, accelerated failure time [Taylor13] *etc.*) where not included in the benchmark, for sake of clear enough plots.

Finally, in our experiments, we did not consider any meta imputation algorithm (post-processor, like GSimp [Wei18, Wang22], ensemble imputation like IMP4P [Giai Gianetto20], chained imputation like MICE [van Buuren11], *etc.*), as their performances directly relates to the imputation algorithm they rely on, and as most of them can be extended to incorporate new input algorithms like Pirat.

3.6.4 Differential abundance validation

On datasets with known ground-truth about differentially expressed proteins [Cox14, Bouyssié20, Huang20b], we compare our methods with those described in Section 3.6.3 using the following

procedure: *(i)* Impute missing peptide (or precursor) abundance values. *(ii)* Test the all-mean equality of each peptide (or precursor) using the one-way ANOVA omnibus test (from R package *stats* [R Core Team13]) and retain the resulting p-values. If a peptide cannot be imputed by a given algorithm (because of too few observed values), the test is performed nonetheless if the number of observed values allows for it (see [R Core Team13] for details), otherwise we set its p-value to one. *(iii)* Display the precision-recall curves for the precision range [90%, 100%] (*i.e.*, low FDP setting) for each method, and compute the global area under the curve.

3.6.5 Mask-and-impute experiments

To precisely evaluate imputation errors, it has become customary to add pseudo-missing values with different MCAR/MNAR proportions, and to compute Root Mean Square Error (RMSE) or Mean Absolute Error (MAE) between imputed and ground-truth values, which read as:

$$\mathrm{RMSE}(X, X^{\mathrm{imp}}) = \sqrt{\frac{1}{n} \sum_{(i,j) \in \mathscr{I}_{\mathrm{pseudo}}} (X_{i,j} - X^{\mathrm{imp}}_{i,j})^2}\,, \tag{3.14}$$

$$\mathrm{MAE}(X, X^{\mathrm{imp}}) = \frac{1}{n} \sum_{(i,j) \in \mathscr{I}_{\mathrm{pseudo}}} |X_{i,j} - X^{\mathrm{imp}}_{i,j}| \tag{3.15}$$

where $\mathscr{I}_{\mathrm{pseudo}}$ is the set of indices of pseudo-missing values and where X^{imp} is the imputed matrix. We generate pseudo MNAR values using a Probit left-censoring mechanism, *i.e.* the probability for a value x to be missing is 1 - $\Phi(x|\nu,\tau)$, where the ν and τ are mean and standard deviation from the Gaussian cumulative distribution function Φ [Miao16]. The overall missing value rate α and the MNAR/MCAR proportion β are controlled by applying the following procedure (adapted from [Lazar16]):

1. Set $\tau = \sigma/2$ where σ is the overall standard deviation of the dataset.
2. For each values of ν' on a linear scale between $\mu - 3\sigma$ and μ (the overall mean of the dataset), compute the expected overall rates q of missing values when applying Probit left-censoring wih parameters ν' and τ, and retain these values. In practice we compute 100 points in total.
3. Interpolate points (ν', q) obtained.
4. Compute ν associated to $\beta\alpha$ (the desired overall MNAR rate) with interpolated curve. If $\alpha\beta$ is not in the range of interpolated curve, choose larger scale for ν' at step 2.
5. Add MNAR values using Probit left-censoring with parameters ν and τ.
6. Add MCAR values with probability $\frac{\alpha(1-\beta)}{1-\alpha\beta}$.

This method ensures that:

- The overall MNAR rate is $\beta\alpha$ (by construction of interpolated curve of (ν', q)).
- The overall MV is rate is α, as

$$\begin{aligned} P(x \text{ is missing}) &= P((x \text{ is MCAR}) \cup (x \text{ is MNAR})) \\ &= P(x \text{ is MCAR}) + P(x \text{ is MNAR}) - P(x \text{ is MCAR} \cap x \text{ is MNAR}) \\ &= \frac{\alpha(1-\beta)}{1-\alpha\beta} + \beta\alpha - \frac{\beta\alpha^2(1-\beta)}{1-\alpha\beta} = \alpha\,, \end{aligned} \tag{3.16}$$

  where x is an observed abundance value.

Finally, we compare all the methods in our mask-and-impute experiments on the very same artificially masked datasets.

3.7 Supplementary Materials

3.7.1 ROC curves

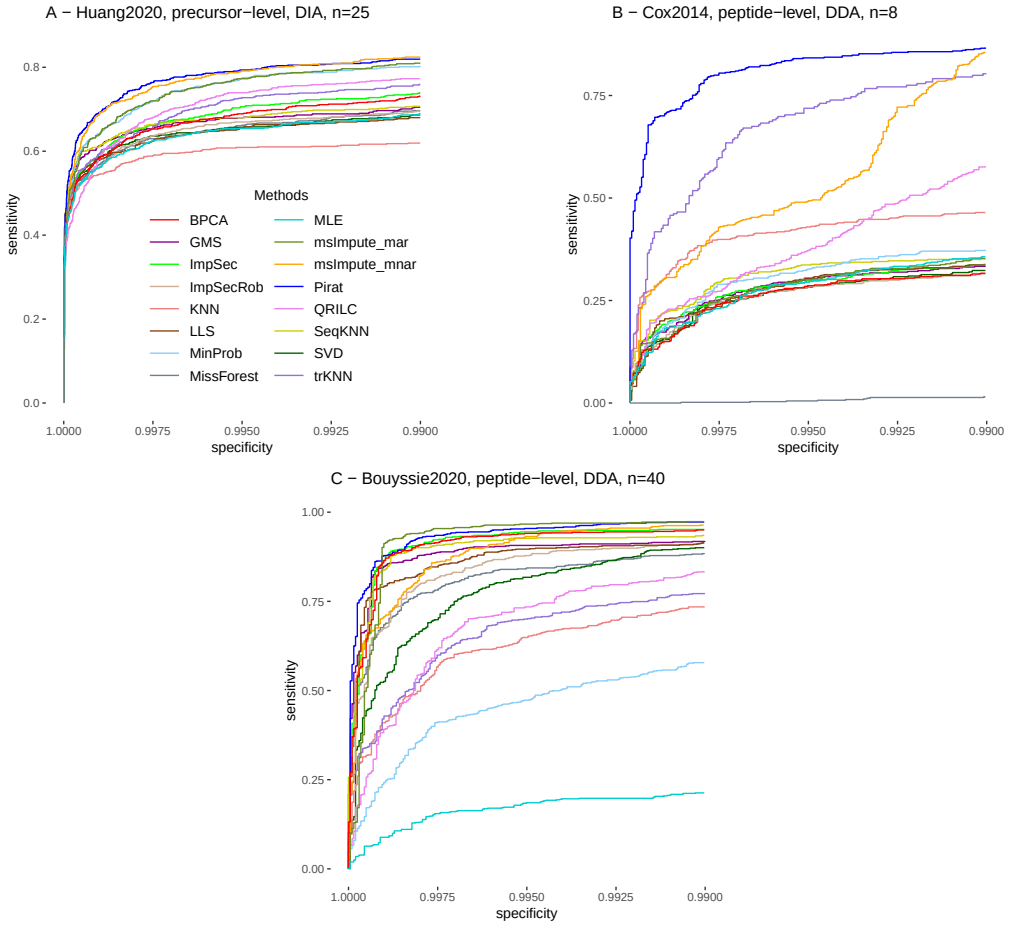

Figure 3.5: ROC curves from p-values associated to the differential analysis validation on the three datasets Huang2020, Cox2014 and Bouyssié2020. We show the curves between 100% and 99% specificity to better differentiate the methods when a stringent selection threshold is applied, which is often the case when FDR is controlled.

3.7.2 Ensembl Gene models, genome assembly, protein databases

This section refers to the filenames of Ensembl gene model, genome assembly and protein databases used to build proteomic and transcriptomic datasets of Ropers2021 and Habowski2020 (see subsection 3.6.2).

Ropers2021

The Ensemble genome assembly used for read alignment in transcriptomic analysis is the following:

`GCA_000005845.2.fasta`

The Ensemble gene model used for mapping slices in transcriptomic analysis and for peptide identification in proteomic alignment is the following :

`Escherichia_coli_str_k_12_substr_mg1655_gca_000005845.ASM584v2.51.gtf`

Habowski2020

The Ensembl genome assembly used for read alignment in transcriptomic analysis is the following:
`Mus_musculus.GRCm39.dna.primary_assembly.fa`

The Ensemble gene model used for mapping slices in transcriptomic analysis is the following:
`Mus_musculus.GRCm39.104.gtf`

The Ensemble protein database used for peptide identification in proteomic analysis is the following:
`Mus_musculus.GRCm39.pep.all.fa`

3.7.3 Mean MAE and RMSE for QRILC and MinProb

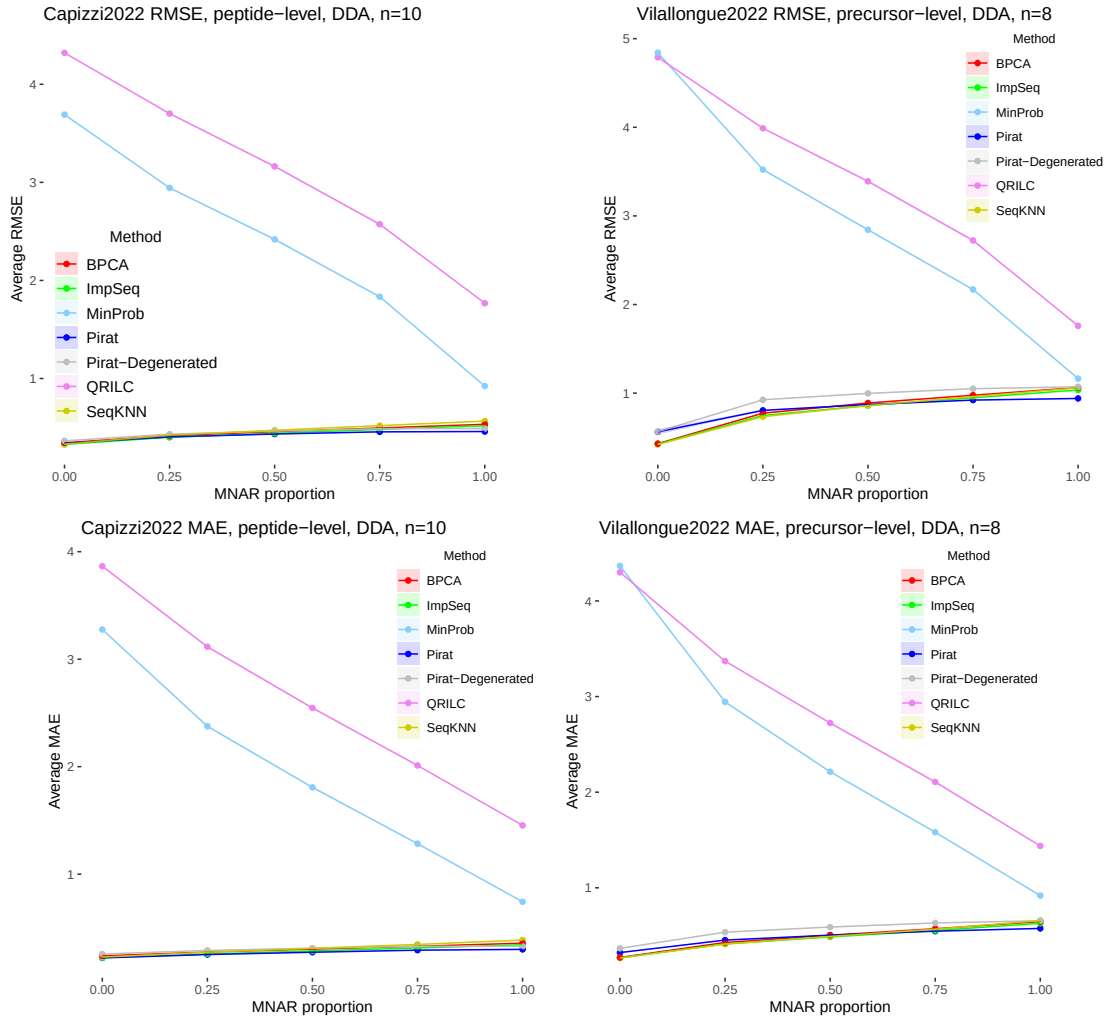

Figure 3.6: Average RMSE (top) and MAE (bottom) of MinProb, QRILC, BPCA, ImpSeq, Pirat, Pirat-Degenerated and SeqKNN in function of the proportion of MNAR values on Capizzi2022 (left) and Villalongue2022 (right). The imputation level (peptide or precursor), the type of acquisition (DIA or DDA), and the total number of replicates (n) are also indicated. The average RMSE and MAE was computed over 5 different seeds, and margins correspond to standard deviation.

3.7.4 Correlations of peptides in Capizzi2022 and Vilallongue2022

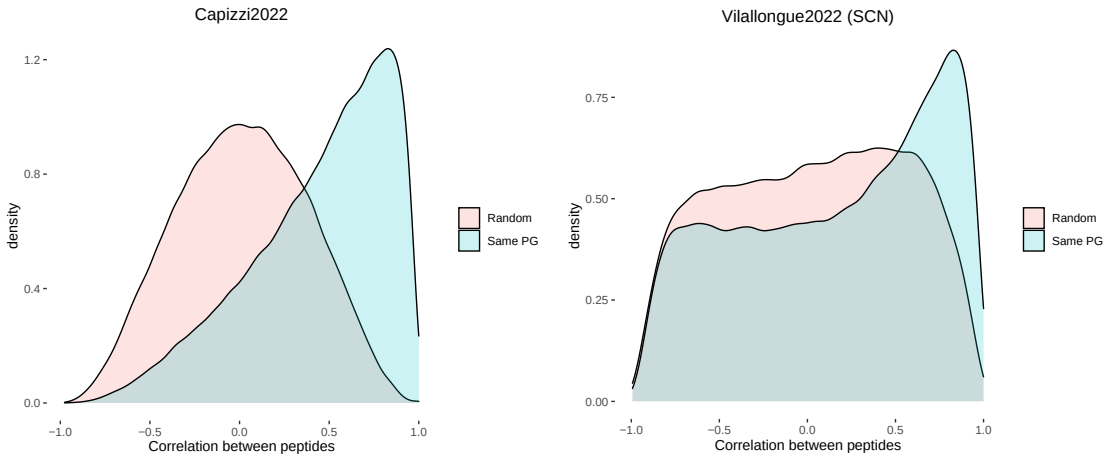

Figure 3.7: Empirical densities of correlations between peptides chosen randomly and between sibling peptides, for Capizzi2022 and Vilallongue2022 (SCN tissue).

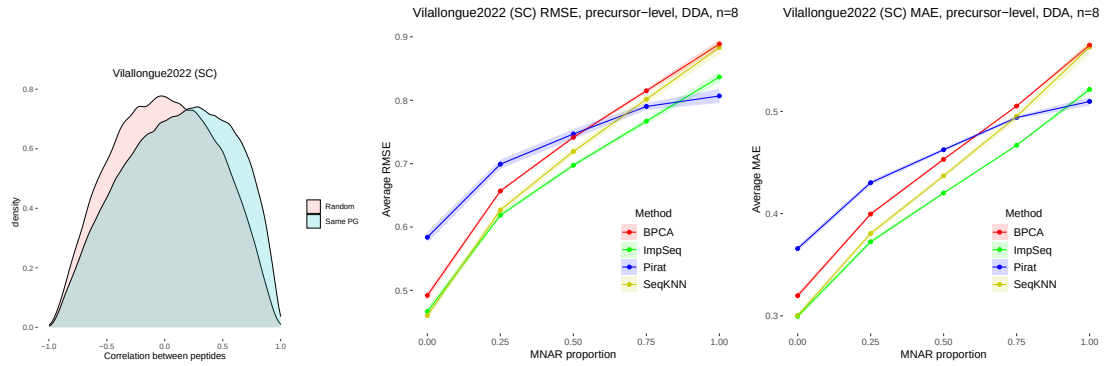

Figure 3.8: Empirical densities of correlations between peptides chosen randomly and between sibling peptides, for Vilallongue2022 (SC tissue), and associated RMSE and MAE vs MNAR proportion curves for different imputation methods on this dataset, according the experimental setting from subsection 3.4.3.

We show on Figure 3.7 the empirical densities of correlations between peptides taken randomly in the dataset, and between siblings, for Capizzi2022 and Vilallongue2022 (the empirical densities are smooth with a gaussian kernel to enhance interpretation). These can be easily plotted with the `plot_pep_correlations` function available in Pirat package. The more the within-PG correlation distribution is right-shifted with respect to that of random correlations, the more within-PGs correlations can be leveraged to impute MVs, and the better Pirat's performances. We observe on Figure 3.7 that the two distributions differ more in Capizzi2022 than in Vilallongue2022, resulting in improved RMSE/MAE performances, especially in MCAR setting (see subsection 3.4.3).

We also give a counter-example of a dataset for which within-PG peptide correlations are very low (the data also come for Vilallongue2022 study, but with a different tissue, referred to as SC) compared to the random peptide correlations (see Figure 3.8). In this setting, Pirat hardly competes with other methods in terms of RMSE and MAE, and only outperforms them in 100% MNAR setting.

3.7.5 Absolute errors for PGs of size one and others on Habowski2020 and Ropers2021

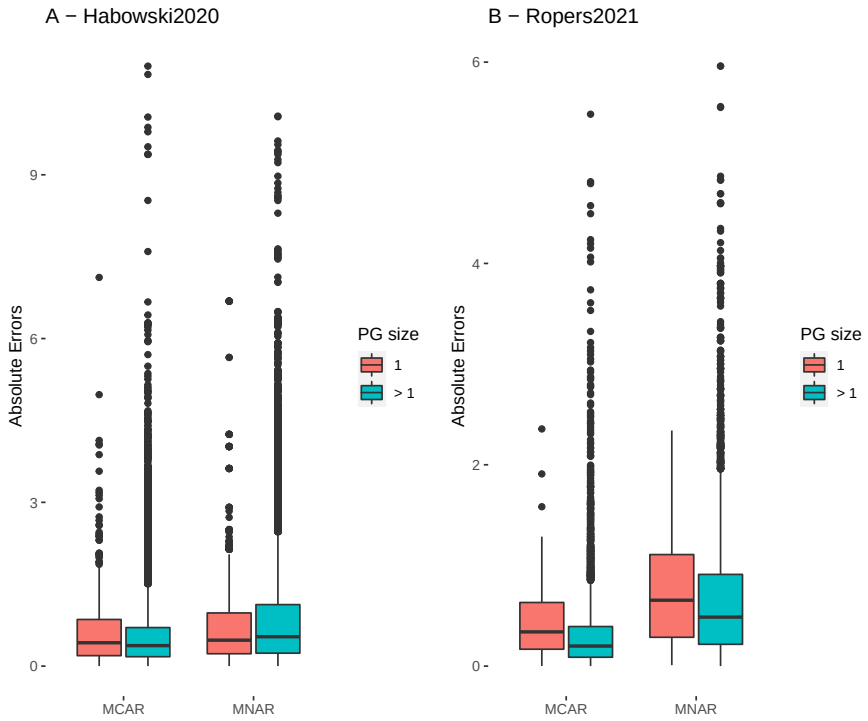

Figure 3.9: Boxplot of absolute errors of Pirat on PG of size equal to one and PGs of size superior to 1 on A - Habowki2020 and B - Ropers2021. Note that in the mask-and-impute experiment, much more pseudo-MVs comes from PGs of size superior to one than in singleton PGs (sometimes 50 times more), which explains the large difference between the number of outliers between for different PG size.

3.7.6 MV distribution in Habowski2020 and Ropers2021 experiments

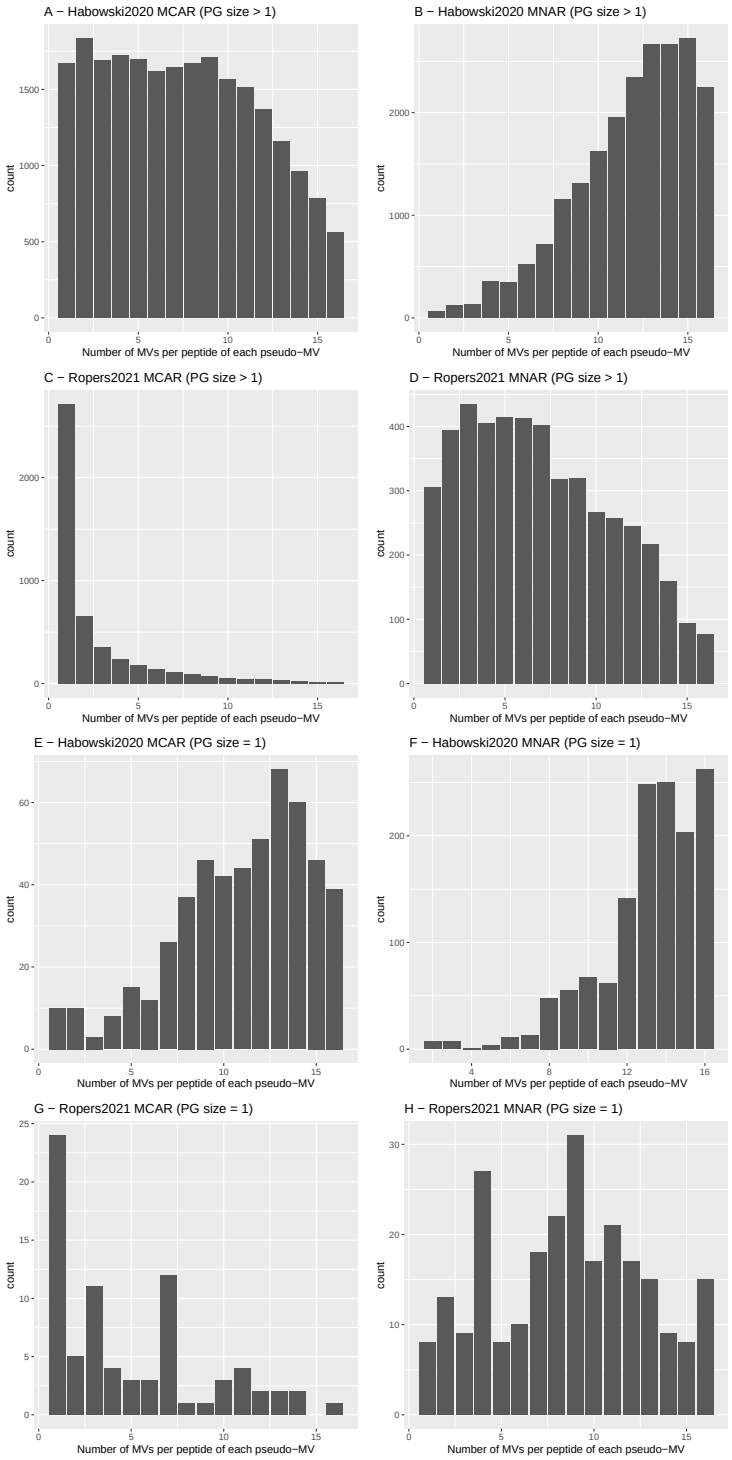

Figure 3.10: Histograms counting, for each pseudo-MV, the number of MVs (real or pseudo) in the same peptide, for Habowski2020 (A, B, E, F) and Ropers2021 (C, D, G, H), in MCAR (A, C, E, G) and MNAR (B, D, F, H) setting, and over 10 different seeds. We separate histograms for peptides contained in singleton PG (A, B, C, D) and non-singleton PGs (E, F, G, H). Note that the number of samples equals 18 in both datasets.

3.7.7 Fitting of missingness mechanism

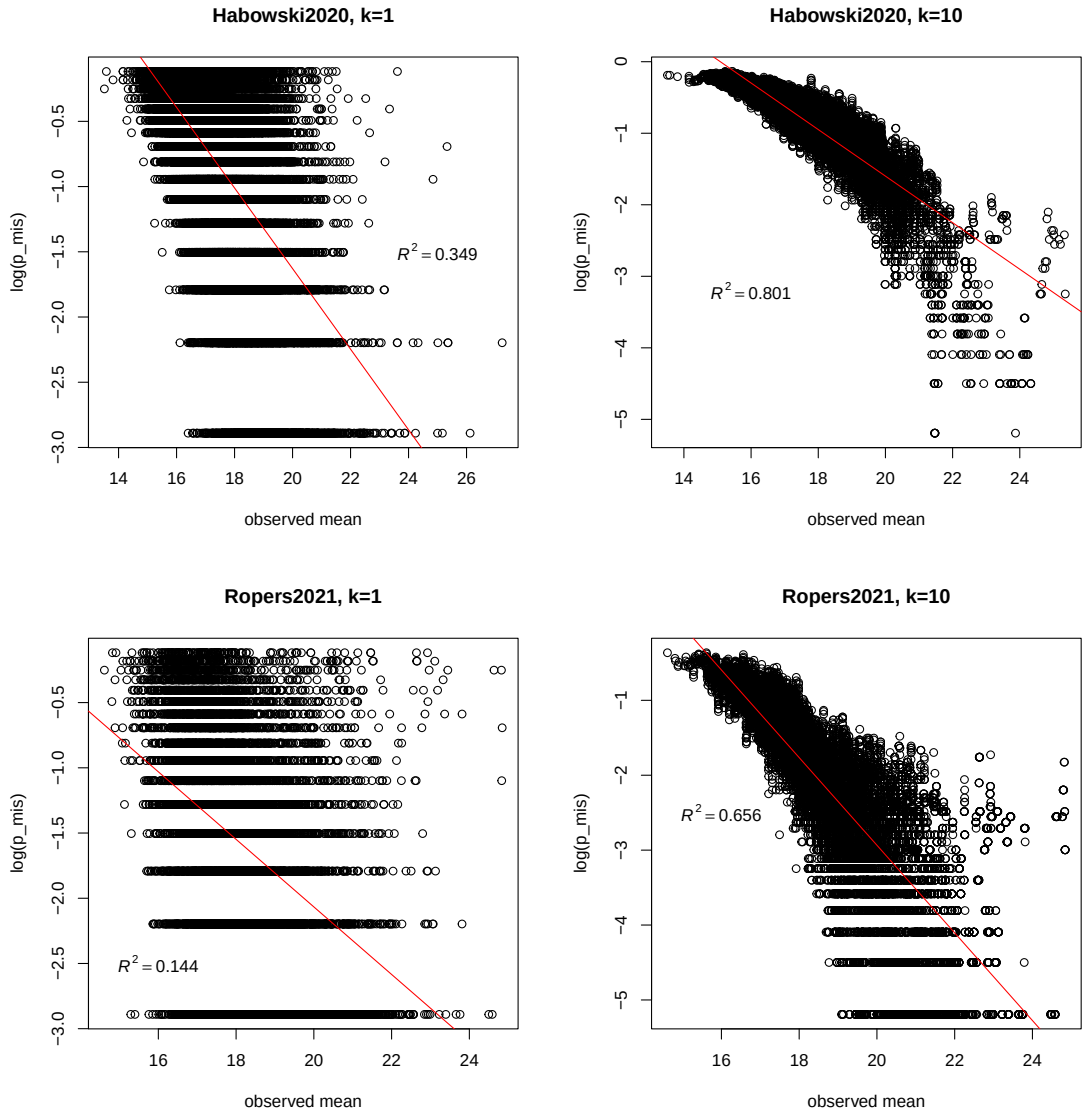

Figure 3.11: Regression of the log-probability of missing onto mean observed abundance following method described in section 3.6.1, for Habowski2020 and Ropers2021, for $k = 1$ and $k = 10$. Residuals sum of squares (R^2) of the linear regression are also displayed.

General conclusions and perspectives

Summary of contributions

My contributions to the statistical treatment of bottom-up label-free LC-MS/MS proteomics data lies at two main levels. First, regarding FDR control in LC-MS/MS proteomics, I investigated the links between empirical methods developed in the proteomics community and theoretical ones used for various purposes: peptide identification and differential analysis. These theoretical considerations enabled us to draw recommendations and guidelines regarding future investigation paths in peptide identification. Second, regarding the MV imputation problem, I developed a novel imputation algorithm, Pirat, which reaches state-of-the-art results on various validation tasks. It is also the first imputation method for LC-MS/MS proteomics allowing for quantitative transcriptomics data integration to help for imputation.

These two research axes both contribute to increasing the peptide coverage and the power of discovery proteomics analyses, while improving their robustness and reliability.

Perspectives

We give here a few perspectives and possible improvements that are directly related to our contributions in chapter 2 and chapter 3, and conclude by a more personal view over the domain.

Differential analysis using knockoffs at peptide level

Our work on imputation suggests alternative ways to generate multivariate knockoffs adapted the high-dimensionality of proteomics data. We concluded in chapter 2 on the original knockoff multivariate approach being not efficient on an entire proteomic dataset, as the estimator of the full protein covariance matrix almost shrinks to a diagonal one. However, at peptide level, the approximation from chapter 3 where within-PG correlations only are considered could be leveraged. This would avoid complete covariance matrix shrinkage, as only a small proportion of covariance weights will be non-zero. We can then expect more powerful knockoff variables, as they would cope for biochemically meaningful dependencies with others, and replicate with greater precision the covariance structure of the original data.

Improving upon Pirat's imputation framework

We discussed at the end of our article (section 3.5) few paths of improvement for Pirat. We provide here additional viewpoints, as we omitted them from the article for the sake of clarity.

First, as depicted on Figure 3.2, Pirat's performances on mask-and-impute experiments are usually better when evaluated using MAE instead of RMSE. This suggests that outlying imputed values decrease the RMSE performances. We noticed that these outliers usually occur when parameters have not converged yet (in particular, the variance parameters remain excessively high).

This could be solved using a more stringent stopping criterion. To avoid a computational time increment on each PG, an adaptive criterion can be sought: For example, requiring that all variances fall below an upper limit derived from the values observed in the dataset. Alternatively, a norm that better copes for outlier values (*e.g.*, the infinity norm) could be interesting.

Second, it would be highly valuable to provide an estimate of the uncertainty of the imputed value. A straightforward way is to compute the variance of the conditional distribution of $X_{i,\text{mis}}$ with respect to all other variables and parameters. It can easily be obtained by Monte-Carlo method, as we can reuse Equation 3.13 to compute the order two moment of the distribution. However, we have observed that this variance estimator has almost no correlation with the actual quadratic errors and is thus a poor predictor of the uncertainty. We assume that this estimator is in fact strongly influenced by the hyperparameter penalty (see section 3.6.1). Whereas this penalty is useful to constrain the estimation and limit aberrant variance values, it may also considerably shrink the variance estimates, and thus the uncertainty of the prediction. Estimation of peptide variances without the penalty could be a remedy, although it would cost additional computational resources. Anyway, a clear advantage of Pirat endowed with such imputation variance estimate, is that we could use it in subsequent analysis with Rubin's rules, for example for differential analysis [Chion22, Chion23], or peptide to protein aggregation.

Improving upon Pirat's transcriptomic integration feature

As mentioned in section 3.1, the transcriptomic integration feature of Pirat has some limitations.

Firstly, the unpaired transcriptomic and proteomic often arises in gene expression studies, and thus Pirat should tackle this issue as best as possible. Our approach (see section 3.6.1) consists in averaging all the transcriptomic intra-condition abundances for each proteomic sample. Although relatively simple, it distorts dramatically the mRNA abundance signal. We have first considered a possible two-step solution that could not be tested because of the constrained timing: first apply Pirat on the intra-condition mean proteomic abundance matrix, as, in unpaired sample setting, only intra-condition means can be paired (therefore, we can impute MEC values by their respective imputed condition mean). Second, impute the POVs with regular Pirat (without transcriptomic integration). This solution avoids the pitfall of variance distortion in transcriptomic data.

Secondly, our proposal for transcriptomic integration only concerns singleton PGs, mainly for coherence and simplification of the article's message. However, we also noticed in both datasets a gain of performances when using Pirat-T on PGs of size 2 and 3. Assessing with a greater precision the PGs that would benefit from transcriptomic integration seems then a natural extension. For example, the package would benefit from a tool that compares the distribution of correlations between transcripts and their associated peptides, with the distribution of intra-PG correlations (similarly to the tool we propose to assess relevance of PGs, see subsection 3.7.4).

Integration of our contributions in the discovery proteomics pipeline

A limit of our contributions, which also applies to many works regarding statistical proteomics, is that the issues we address are often considered independently from the rest of the analysis pipeline. Indeed, because of its overall complexity, we cannot assess the combination of every concurrent method for each step of the data processing workflow (FDR control, normalization, imputation, filtering, *etc.*). However, because of some tight dependencies between some of these steps, studying them in a more holistic manner would be beneficial to the field.

For example, the Match-Between-Run option (MBR, see section 1.1.4) raises major issues on the data treatment, as is can affect more than 20% of total abundances produced in typical experiments at EDyP Lab. First, MBR completely bypasses the FDR control procedure at peptide identification. This is a long running issue and it will be investigated in a follow up project in the lab. Second, as MBR and statistical imputation essentially play the same roles, *i.e.*, completing the dataset, it would be interesting to assess them together. For example, one might be preferable to

the other depending on various factors (MECs, POVS, type of dataset, *etc.*) and, to the best of our knowledge, no comparative study has been published on the subject. Yet, some preliminary tests in EDyP have shown MBR to be more accurate than imputation on a benchmark dataset, but more exhaustive and robust assessments are required.

A second important link to consider is the relationship between imputation and differential analysis. We have partly addressed this issue in chapter 3 by comparing the imputation methods on differential analysis task. Yet, we can go further by comparing Pirat with dedicated imputation-free differential analysis tools (see subsection 1.3.2). If Pirat achieves similar (or better) performances than imputation free methods, it then proves Pirat could be used in routine regardless of the downstream analysis. Another related subject pertains to the influence of imputation on the overall distribution of the final p-values, and thus on the quality of FDR control when using BH type procedures. A convenient imputation method should not alter the distribution of p-values of peptides under the null hypothesis, which should remain close to the uniform one (supposing correct calibration before imputation). This evaluation metric has been used for example in msImpute [Hediyeh-zadeh23]. Yet, uniformity of p-values under the null hypothesis is not mandatory to control FDR. Notably, as we have seen in section 1.2, knockoffs filters with p-value based score (see section 2.2) can overcome the uniformity limitation.

Finally, a last connection of our work with the overall data analysis pipeline pertains to the question of imputing before or after precursor-to-peptide aggregation. Although we do not give conclusion about this topic in chapter 3 (as we did not make any experiments to support this claim), I think it is preferable to impute before aggregation to avoid an implicit imputation, as advocated in favor of imputing before peptide-to-protein aggregation. However, another important argument is that the estimation of the missingness mechanism should directly relate to the stochastic limit of detection of the instrument, and hence should be involved as close as possible from the instrument's output, *i.e.*, on precursors.

Personal view on missing value imputation in gene expression data

The abundant literature on imputation-free methods (see subsection 1.3.2), as well as the predominance of differential analysis in proteomics experiments made me wonder about the practical interest of imputation during my bibliographical research. My perspective on this has somewhat shifted since then.

Precisely, I don't think imputation is a sensible approach when there is no helpful external information, and thus agree with Chion *et al.* [Chion23]. However, in LC-MS/MS data, it is doubly not the case. First, as our work suggests, there are known biochemical dependencies to exploit for missing data inference. Second, in the presence of an abundance dependent missingness mechanism, the fact that the value is missing actually brings information in counterpart. In recent work, Chion *et al.* [Chion23] only consider the first point and clearly argues against imputation in univariate setting. However, they overlook the second point, which, in my opinion, is fundamental and adds subtlety to their perspective. For example, if we have a known fixed detection threshold, then a univariate imputation by a value slightly lower than this detection threshold makes more sense than relying solely on variable dependencies. In the specific case of LC-MS/MS proteomics data, opting for univariate imputation can be a sensible choice, as we can always rely on the presence of the left-censoring mechanism. The univariate version of Pirat (that we refer to as "Pirat-Degenerated") actually competes with the other MAR multivariate methods in full MNAR setting (see Figure 3.2). However, in order to rely on the presence of such left-censoring mechanism, it has to be "brutal" enough to be informative. By "brutal", I mean here that $P(M_{i,j} = 1|X_{i,j})$ should significantly decrease when $X_{i,j}$ increases (*i.e.*, in Pirat framework, true γ_1 should be high enough). Otherwise, the missingness indicator is poorly related to the abundance value, and relying only on this in a univariate imputation method would be suboptimal.

Finally, a last argument towards the development of robust imputation algorithms is that it

encourages the development other types of downstream analysis, differing from differential analysis, and thus widen the possible usages of LC-MS/MS proteomics analysis.

Personal overview and general perspectives on discovery proteomics for biomedical research

At the beginning of my thesis, the high dimensionality of the data in bulk proteomics experiments (few samples for tens of thousands of peptides) seemed a major challenge to me. In my opinion, many multivariate analysis tools used in areas such as multi-omics integration, along with knockoff filters, were not tailored to accommodate the curse of the dimensionality of the data. Finding new alternatives appeared challenging. However, at least regarding imputation, I realized that the PG strategy had never been proposed yet and could be a natural way to tackle this dimensionality issue.

Yet the number of samples obtained in bulk discovery proteomics studies remains, according to me, a major limitation to innovation in the statistical treatment of the produced data. For example, in most studies, we are essentially constrained to perform univariate differential analysis. Although this procedure is very efficient to discover a few but important biomarkers, often endowed with meaningful biological interpretation, it is limiting when the objective is to build a composite biomarker, as correlations are not accounted for (see section 2.5). For example, the proteomic signature of the development of a disease may include a large amount of proteins, thus requiring huge cohorts to model their interactions (*e.g.* $n > 500$ in a recent study [Niu22]), which are rather rare because of the cost of MS analyzes. On top of that, and until now, even if we obtain larger sample size, we have weak guarantees that our data processing methods (aggregation, imputation, *etc.*) would be well suited to the corresponding applications, as most of validation procedures are oriented towards differential analysis.

Through conversations with a data scientist from OWKIN, a biomedical research company, and talks given by researchers at ISMB EECB 2023 conference, I noticed that LC-MS/MS proteomics analyses were rarely considered, probably for the latter reasons. Oppositely, and according to them, single-cell approaches are gaining more and more interest in biomedical research. These are particularly showcased for cancer research, to tackle the enormous variability among the cancerous cells. Some scRNA-seq technologies enable to analyse thousands of cells per run and can include additional spatial information. Hence, these analyses do not aim at discovering one or few biomarkers.

Instead, large multivariate models can build latent representations of cells and tissues, which can then be employed for a various tasks, such as clustering or classifying cancerous cells, predicting their evolution, *etc.* Yet, the processing of scRNA-seq data, as of LC-MS/MS proteomics, is not exempt from limitations: a great amount of zero expressed transcript with complex origins [Lindermanan22], normalization issues between samples [Stuart19], dubious cell embedding [Xia23], overall lack of validation and quality control procedures, *etc.* Naturally, these problems are not unlike those encountered in this manuscript. Besides the obstacles related to single-cell data analysis, the analysis of scMS proteomics data encounters challenges similar to those encountered in classical "bulk" bottom-up approaches. However, this technology is more recent than scRNA-seq, and neither its gene covering, or cell throughput is comparable yet to that of scRNA-seq [Bennett23]. Overall, the need for robust and reliable statistical treatment in single-cell analyses (both scMS and scRNA-seq), along with the wide possibilities offered by the sample size, motivates me to consider this domain in the rest of my carrier.

Bibliography

[Alberts17] B. Alberts. *Molecular biology of the cell*. Garland science, 2017.

[Anders15] S. Anders, P. T. Pyl, and W. Huber. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTU638.

[Anderson98] N. L. Anderson and N. G. Anderson. Proteome and proteomics: New technologies, new concepts, and new words. *ELECTROPHORESIS*, 19(11):1853–1861, 1998. ISSN 1522-2683. doi:10.1002/ELPS.1150191103.

[Argelaguet18] R. Argelaguet, B. Velten, D. Arnol, *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, 2018. ISSN 1744-4292. doi:10.15252/msb.20178124.

[Bantscheff12] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry 2012 404:4*, 404(4):939–965, 2012. ISSN 1618-2650. doi:10.1007/S00216-012-6203-4.

[Barber15] R. F. Barber and E. J. Candés. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015. ISSN 00905364. doi:10.1214/15-AOS1337.

[Barzine20] M. P. Barzine, K. Freivalds, J. C. Wright, *et al.* Using Deep Learning to Extrapolate Protein Expression Measurements. *Proteomics*, 20(21-22), 2020. ISSN 16159861. doi:10.1002/pmic.202000009.

[Bateman14] N. W. Bateman, S. P. Goulding, N. J. Shulman, *et al.* Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (DDA). *Molecular and Cellular Proteomics*, 13(1):329–338, 2014. ISSN 15359476. doi:10.1074/mcp.M112.026500.

[Bateman21] A. Bateman, M. J. Martin, S. Orchard, *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2021. ISSN 0305-1048. doi:10.1093/NAR/GKAA1100.

[Behjati13] S. Behjati and P. S. Tarpey. What is next generation sequencing? *Archives of Disease in Childhood. Education and Practice Edition*, 98(6):236, 2013. ISSN 17430585. doi:10.1136/ARCHDISCHILD-2013-304340.

[Benjamini95] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. ISSN 2517-6161. doi:10.1111/J.2517-6161.1995.TB02031.X.

[Benjamini01] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. ISSN 0090-5364. doi:10.1214/AOS/1013699998.

[Benjamini06] Y. Benjamini, A. M. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006. doi:10.1093/biomet/93.3.491.

[Benjamini10] Y. Benjamini. Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):405–416, 2010. ISSN 1467-9868. doi:10.1111/J.1467-9868.2010.00746.X.

[Benjamini16] Y. Benjamini and Y. Hochberg. On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics:. *Journal of Educational and Behavioral Statistics*, 25(1):60–83, 2016. ISSN 10769986. doi:10.3102/10769986025001060.

[Bennett23] H. M. Bennett, W. Stephenson, C. M. Rose, and S. Darmanis. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nature Methods 2023 20:3*, 20(3):363–374, 2023. ISSN 1548-7105. doi:10.1038/s41592-023-01791-5.

[Blein-Nicolas16] M. Blein-Nicolas and M. Zivy. Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1864(8):883–895, 2016. ISSN 1570-9639. doi:10.1016/J.BBAPAP.2016.02.019.

[Bolger14] A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTU170.

[Bouret18] P. Bouret and F. Bastien. Erreurs et tests statistiques (40 min), 2018.

[Bouyssié20] D. Bouyssié, A. M. Hesse, E. Mouton-Barbosa, *et al.* Proline: an efficient and user-friendly software suite for large-scale proteomics. *Bioinformatics*, 36(10):3148–3155, 2020. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTAA118.

[Branden09] K. V. Branden and S. Verboven. Robust data imputation. *Computational Biology and Chemistry*, 33(1):7–13, 2009. ISSN 1476-9271. doi:10.1016/J.COMPBIOLCHEM.2008.07.019.

[Bubis17] J. A. Bubis, L. I. Levitsky, M. V. Ivanov, *et al.* Comparative evaluation of label-free quantification methods for shotgun proteomics. *Rapid Communications in Mass Spectrometry*, 31(7):606–612, 2017. ISSN 1097-0231. doi:10.1002/RCM.7829.

[Burger18] T. Burger. Gentle Introduction to the Statistical Foundations of False Discovery Rate in Quantitative Proteomics. *Journal of Proteome Research*, 17(1):12–22, 2018. ISSN 15353907. doi:10.1021/ACS.JPROTEOME.7B00170.

[Candès18] E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018. ISSN 1467-9868. doi:10.1111/RSSB.12265.

[Capizzi22] M. Capizzi, R. Carpentier, E. Denarier, *et al.* Developmental defects in Huntington's disease show that axonal growth and microtubule reorganization require NUMA1. *Neuron*, 110(1):36–50.e5, 2022. ISSN 0896-6273. doi:10.1016/J.NEURON.2021.10.033.

[Carpenter00] J. Carpenter and J. Bithell. Bootstrap con"dence intervals: when, which, what? A practical guide for medical statisticians. *STATISTICS IN MEDICINE Statist. Med*, 19:1141–1164, 2000. doi:10.1002/(SICI)1097-0258(20000515)19:9.

[Chen14] L. S. Chen, R. L. Prentice, and P. Wang. A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. *Biometrics*, 70(2):312–322, 2014. ISSN 1541-0420. doi:10.1111/BIOM.12149.

[Chion21] M. Chion. *Development of new statistical methodologies for quantitative proteomics data analysis*. Ph.D. thesis, Université de Strasbourg, 2021. doi:10.13039/501100001665.

[Chion22] M. Chion, C. Carapito, and F. Bertrand. Accounting for multiple imputation-induced variability for differential analysis in mass spectrometry-based label-free quantitative proteomics. *PLOS Computational Biology*, 18(8):e1010420, 2022. ISSN 1553-7358. doi:10.1371/journal.pcbi.1010420.

[Chion23] M. Chion and A. Leroy. A bayesian framework for multivariate differential analysis accounting for missing data, 2023. doi:10.48550/arXiv.2307.08975.

[Choi08] H. Choi and A. I. Nesvizhskii. False Discovery Rates and Related Statistical Concepts in Mass Spectrometry-Based Proteomics. *Journal of Proteome Research*, 7(1):47–50, 2008. ISSN 1535-3893. doi:10.1021/pr700747q.

[Cobb17] M. Cobb. 60 years ago, Francis Crick changed the logic of biology. *PLOS Biology*, 15(9):e2003243, 2017. ISSN 1545-7885. doi:10.1371/JOURNAL.PBIO.2003243.

[Colinge03] J. Colinge, A. Masselot, M. Giron, *et al.* OLAV: Towards high-throughput tandem mass spectrometry data identification. *PROTEOMICS*, 3(8):1454–1463, 2003. ISSN 1615-9861. doi:10.1002/PMIC.200300485.

[Cooper12] B. Cooper. The Problem with Peptide Presumption and the Downfall of Target–Decoy False Discovery Rates. *Analytical Chemistry*, 84(22):9663–9667, 2012. ISSN 0003-2700. doi:10.1021/ac303051s.

[Couté20] Y. Couté, C. Bruley, and T. Burger. Beyond Target–Decoy Competition: Stable Validation of Peptide and Protein Identifications in Mass Spectrometry-Based Discovery Proteomics. *Analytical Chemistry*, 92(22):14898–14906, 2020. ISSN 0003-2700. doi:10.1021/acs.analchem.0c00328.

[Cox11] J. Cox, N. Neuhauser, A. Michalski, *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, 10(4):1794–1805, 2011. ISSN 15353893. doi:10.1021/PR101065J.

[Cox14] J. Cox, M. Y. Hein, C. A. Luber, *et al.* Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics : MCP*, 13(9):2513, 2014. ISSN 15359484. doi:10.1074/MCP.M113.031591.

[Craig04] R. Craig and R. C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004. ISSN 1367-4811. doi:10.1093/bioinformatics/bth092.

[DeLong88] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3):837, 1988. ISSN 0006341X. doi:10.2307/2531595.

[Dermit21] M. Dermit and J. G. Meyer. Peptide Correlation Analysis (PeCorA) Reveals Differential Proteoform Regulation. *Journal of Proteome Research*, 2021. ISSN 15353907. doi:10.1021/acs.jproteome.0c00602.

[Desaire22] H. Desaire. How (Not) to Generate a Highly Predictive Biomarker Panel Using Machine Learning. *Journal of Proteome Research*, 21(9):2071–2074, 2022. ISSN 15353907. doi:10.1021/ACS.JPROTEOME.2C00117.

[Dobin13] A. Dobin, C. A. Davis, F. Schlesinger, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTS635.

[Doerr14] A. Doerr. DIA mass spectrometry. *Nature Methods 2015 12:1*, 12(1):35–35, 2014. ISSN 1548-7105. doi:10.1038/nmeth.3234.

[Dulai17] P. S. Dulai, S. Singh, J. Patel, *et al.* Increased risk of mortality by fibrosis stage in nonalcoholic fatty liver disease: Systematic review and meta-analysis. *Hepatology*, 65(5):1557–1565, 2017. ISSN 15273350. doi:10.1002/HEP.29085.

[Edfors16] F. Edfors, F. Danielsson, B. M. Hallström, *et al.* Gene-specific correlation of RNA and protein levels in human cells and tissues . *Molecular Systems Biology*, 12(10):883, 2016. ISSN 1744-4292. doi:10.15252/msb.20167144.

[Efron72] B. Efron and C. Morris. Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case. *Journal of the American Statistical Association*, 67(337):130–139, 1972. ISSN 0162-1459. doi:10.1080/01621459.1972.10481215.

[Efron01] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001. ISSN 1537274X. doi:10.1198/016214501753382129.

[Efron02] B. Efron and R. Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86, 2002. ISSN 1098-2272. doi:10.1002/GEPI.1124.

[Efron04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2), 2004. ISSN 0090-5364. doi:10.1214/009053604000000067.

[Elias07a] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods 2007 4:3*, 4(3):207–214, 2007. ISSN 1548-7105. doi:10.1038/nmeth1019.

[Elias07b] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods 2007 4:3*, 4(3):207–214, 2007. ISSN 1548-7105. doi:10.1038/nmeth1019.

[Emery19] K. Emery, S. Hasam, W. S. Noble, and U. Keich. Multiple competition-based FDR control for peptide detection. *arXiv*, pages 1–34, 2019. doi:10.48550/arXiv.1907.01458.

[Emery20] K. Emery, S. Hasam, W. S. Noble, and U. Keich. Multiple Competition-Based FDR Control and Its Application to Peptide Detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12074 LNBI:54–71, 2020. ISSN 16113349. doi:10.1007/978-3-030-45257-5_4.

[Etourneau22] L. Etourneau and T. Burger. Challenging Targets or Describing Mismatches? A Comment on Common Decoy Distribution by Madej et al. *Journal of Proteome Research*, 21(12):2840–2845, 2022. ISSN 15353907. doi:10.1021/ACS.JPROTEOME.2C00279.

[Etourneau23] L. Etourneau, N. Varoquaux, and T. Burger. Unveiling the Links Between Peptide Identification and Differential Analysis FDR Controls by Means of a Practical Introduction to Knockoff Filters. *Methods in Molecular Biology*, 2426:1–24, 2023. ISSN 19406029. doi:10.1007/978-1-0716-1967-4_1.

[Fancello22] L. Fancello and T. Burger. An analysis of proteogenomics and how and when transcriptome-informed reduction of protein databases can enhance eukaryotic proteomics. *Genome Biology*, 23(1):1–23, 2022. ISSN 1474760X. doi:10.1186/S13059-022-02701-2.

[Flores23] J. E. Flores, D. M. Claborne, Z. D. Weller, *et al.* Missing data in multi-omics integration: Recent advances through artificial intelligence. *Frontiers in Artificial Intelligence*, 6, 2023. ISSN 26248212. doi:10.3389/frai.2023.1098308.

[Fortelny17] N. Fortelny, C. M. Overall, P. Pavlidis, and G. V. C. Freue. Can we predict protein from mRNA levels? *Nature 2017 547:7664*, 547(7664):E19–E20, 2017. ISSN 1476-4687. doi:10.1038/nature22293.

[Friedman10] J. Friedman, J. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi:10.18637/jss.v033.i01.

[Fröhlich22] K. Fröhlich, E. Brombacher, M. Fahrner, *et al.* Benchmarking of analysis strategies for data-independent acquisition proteomics using a large-scale dataset comprising inter-patient heterogeneity. *Nature Communications 2022 13:1*, 13(1):1–13, 2022. ISSN 2041-1723. doi:10.1038/s41467-022-30094-0.

[Gadiparthi20] C. Gadiparthi, M. Spatz, S. Greenberg, *et al.* NAFLD Epidemiology, Emerging Pharmacotherapy, Liver Transplantation Implications and the Trends in the United States. *Journal of Clinical and Translational Hepatology*, 8(2):215, 2020. ISSN 23108819. doi:10.14218/JCTH.2020.00014.

[Gardner21] M. L. Gardner and M. A. Freitas. Multiple Imputation Approaches Applied to the Missing Value Problem in Bottom-Up Proteomics. *International Journal of Molecular Sciences*, 22(17):9650, 2021. ISSN 1422-0067. doi:10.3390/ijms22179650.

[Ge03] Y. Ge, S. Dudoit, and T. P. Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77, 2003. ISSN 0041-0241. doi:10.1007/BF02595811.

[Ge21] X. Ge, Y. E. Chen, D. Song, *et al.* Clipper: p-value-free FDR control on high-throughput data from two conditions. *Genome Biology*, 22(1):288, 2021. ISSN 1474-760X. doi:10.1186/s13059-021-02506-9.

[Geer04] L. Y. Geer, S. P. Markey, J. A. Kowalak, *et al.* Open mass spectrometry search algorithm. *Journal of Proteome Research*, 3(5):958–964, 2004. ISSN 15353893. doi:10.1021/PR0499491.

[Giai Gianetto16] Q. Giai Gianetto, F. Combes, C. Ramus, *et al.* Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. *PROTEOMICS*, 16(1):29–32, 2016. ISSN 1615-9861. doi:10.1002/PMIC.201500189.

[Giai Gianetto19] Q. Giai Gianetto, F. Combes, C. Ramus, *et al. cp4p: Calibration Plot for Proteomics*, 2019. R package version 0.3.6.

[Giai Gianetto20] Q. Giai Gianetto, S. Wieczorek, Y. Couté, and T. Burger. A peptide-level multiple imputation strategy accounting for the different natures of missing values in proteomics data. *bioRxiv*, page 2020.05.29.122770, 2020. doi:10.1101/2020.05.29.122770.

[Gillet16] L. C. Gillet, A. Leitner, and R. Aebersold. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *https://doi.org/10.1146/annurev-anchem-071015-041535*, 9:449–472, 2016. ISSN 19361335. doi:10.1146/ANNUREV-ANCHEM-071015-041535.

[Gimenez18] J. R. Gimenez and J. Zou. Improving the Stability of the Knockoff Procedure: Multiple Simultaneous Knockoffs and Entropy Maximization. *AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics*, 89, 2018. doi:10.48550/arxiv.1810.11378.

[Goeman14] J. J. Goeman and A. Solari. Multiple hypothesis testing in genomics. *Statistics in Medicine*, 33(11):1946–1978, 2014. ISSN 1097-0258. doi:10.1002/SIM.6082.

[Goeminne20] L. J. E. Goeminne, A. Sticker, L. Martens, *et al.* MSqRob Takes the Missing Hurdle: Uniting Intensity- and Count-Based Proteomics. *Analytical Chemistry*, 92(9):6278–6287, 2020. ISSN 0003-2700. doi:10.1021/acs.analchem.9b04375.

[Gupta23] S. Gupta, C. N. Schlaffner, S. Ahmed, *et al.* Imputing protein abundance by modeling molecular relationships using gnns. `https://www.iscb.org/ismbeccb2023-programme/posters`, 2023. Accessed: 2023-10-05.

[Habowski20] A. N. Habowski, J. L. Flesher, J. M. Bates, *et al.* Transcriptomic and proteomic signatures of stemness and differentiation in the colon crypt. *Communications Biology 2020 3:1*, 3(1):1–17, 2020. ISSN 2399-3642. doi:10.1038/s42003-020-01181-z.

[Harris23] L. Harris, W. E. Fondrie, S. Oh, and W. S. Noble. Evaluating Proteomics Imputation Methods with Improved Criteria. *Journal of Proteome Research*, 2023. ISSN 1535-3893. doi:10.1021/ACS.JPROTEOME.3C00205.

[Hastie13] T. Hastie and B. Efron. *LARS: Least Angle Regression, Lasso and Forward Stagewise*, 2013. R package version 1.2.

[He15] K. He, Y. Fu, W.-F. Zeng, *et al.* A theoretical foundation of the target-decoy search strategy for false discovery rate control in proteomics. *arXiv*, 2015. doi:10.48550/arxiv.1501.00537.

[He18a] K. He, M. Li, Y. Fu, *et al.* Null-free False Discovery Rate Control Using Decoy Permutations. *Acta Mathematicae Applicatae Sinica, English Series 2022 38:2*, 38(2):235–253, 2018. ISSN 16183932. doi:10.1007/s10255-022-1077-5.

[He18b] K. He, M. Li, Y. Fu, *et al.* Null-free False Discovery Rate Control Using Decoy Permutations. *Acta Mathematicae Applicatae Sinica, English Series 2022 38:2*, 38(2):235–253, 2018. ISSN 16183932. doi:10.1007/s10255-022-1077-5.

[Hediyeh-zadeh23] S. Hediyeh-zadeh, A. I. Webb, and M. J. Davis. MsImpute: Estimation of missing peptide intensity data in label-free quantitative mass spectrometry. *Molecular & Cellular Proteomics*, page 100558, 2023. ISSN 1535-9476. doi:10.1016/J.MCPRO.2023.100558.

[Hinz10] U. Hinz and T. U. Consortium. From protein sequences to 3D-structures and beyond: the example of the UniProt Knowledgebase. *Cellular and Molecular Life Sciences*, 67(7):1049, 2010. ISSN 1420682X. doi:10.1007/S00018-009-0229-6.

[Huang20a] D. Q. Huang, H. B. El-Serag, and R. Loomba. Global epidemiology of NAFLD-related HCC: trends, predictions, risk factors and prevention. *Nature Reviews Gastroenterology & Hepatology 2020 18:4*, 18(4):223–238, 2020. ISSN 1759-5053. doi:10.1038/s41575-020-00381-6.

[Huang20b] T. Huang, R. Bruderer, J. Muntel, *et al.* Combining Precursor and Fragment Information for Improved Detection of Differential Abundance in Data Independent Acquisition. *Molecular & Cellular Proteomics*, 19(2):421–430, 2020. ISSN 1535-9476. doi:10.1074/MCP.RA119.001705.

[Huber03] W. Huber, A. von Heydebreck, H. Sueltmann, *et al.* Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2(1), 2003. ISSN 1544-6115. doi:10.2202/1544-6115.1008.

[Huber15] W. Huber, V. J. Carey, R. Gentleman, *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, 2015. doi:10.1038/nmeth.3252.

[Ilin10] A. Ilin and T. Raiko. Practical Approaches to Principal Component Analysis in the Presence of Missing Values. *Journal of Machine Learning Research*, 11:1957–2000, 2010. doi:10.5555/1756006.1859917.

[Jacob16] L. Jacob, J. A. Gagnon-Bartsch, and T. P. Speed. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*, 17(1):16–28, 2016. ISSN 1465-4644. doi:10.1093/BIOSTATISTICS/KXV026.

[Jeong12] K. Jeong, S. Kim, and N. Bandeira. False discovery rates in spectral identification. *BMC bioinformatics*, 13 Suppl 1(16):1–15, 2012. ISSN 14712105. doi:10.1186/1471-2105-13-S16-S2.

[Jin21] L. Jin, Y. Bi, C. Hu, *et al.* A comparative study of evaluating missing value imputation methods in label-free proteomics. *Scientific Reports*, 11(1):1760, 2021. ISSN 20452322. doi:10.1038/s41598-021-81279-4.

[Josse18] J. Josse. Handling missing values. `https://juliejosse.com/wp-content/uploads/2018/07/LectureNotesMissing.html#2)_ml_inference_with_missing_values`, 2018. Accessed: 2023-10-25.

[Käll07] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *Journal of Proteome Research*, 7(1):29–34, 2007. ISSN 15353893. doi:10.1021/PR700600N.

[Käll08] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of Proteome Research*, 7(1):29–34, 2008. ISSN 15353893. doi:10.1021/pr700600n.

[Kanter15] I. Kanter and T. Kalisky. Single Cell Transcriptomics: Methods and Applications. *Frontiers in Oncology*, 5(FEB), 2015. ISSN 2234943X. doi:10.3389/FONC.2015.00053.

[Karpievitch09] Y. Karpievitch, J. Stanley, T. Taverner, *et al.* A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics*, 25(16):2028–2034, 2009. ISSN 13674803. doi:10.1093/bioinformatics/btp362.

[Karpievitch12] Y. V. Karpievitch, A. R. Dabney, and R. D. Smith. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics 2012 13:16*, 13(16):1–9, 2012. ISSN 1471-2105. doi:10.1186/1471-2105-13-S16-S5.

[Keich15] U. Keich, A. Kertesz-Farkas, and W. S. Noble. Improved False Discovery Rate Estimation Procedure for Shotgun Proteomics. *Journal of Proteome Research*, 14(8):3148–3161, 2015. ISSN 15353907. doi:10.1021/acs.jproteome.5b00081.

[Keich19] U. Keich, K. Tamura, and W. S. Noble. Averaging Strategy To Reduce Variability in Target-Decoy Estimates of False Discovery Rate. *Journal of Proteome Research*, 18(2):585–593, 2019. ISSN 15353907. doi:10.1021/ACS.JPROTEOME.8B00802.

[Keller02] A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Analytical Chemistry*, 74(20):5383–5392, 2002. ISSN 0003-2700. doi:10.1021/ac025747h.

[Kim04] K. Y. Kim, B. J. Kim, and G. S. Yi. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics*, 5(1):1–9, 2004. ISSN 14712105. doi:10.1186/1471-2105-5-160.

[Kim05] H. Kim, G. H. Golub, and H. Park. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198, 2005. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTH499.

[Klammer06] A. A. Klammer and M. J. MacCoss. Effects of modified digestion schemes on the identification of proteins from complex mixtures. *Journal of Proteome Research*, 5(3):695–700, 2006. ISSN 15353893. doi:10.1021/PR050315J.

[Kohler23] D. Kohler, M. Staniak, T. H. Tsai, *et al.* MSstats Version 4.0: Statistical Analyses of Quantitative Mass Spectrometry-Based Proteomic Experiments with Chromatography-Based Quantification at Scale. *Journal of Proteome Research*, 22(5):1466–1482, 2023. ISSN 15353907. doi:10.1021/ACS.JPROTEOME.2C00834.

[Kong22] W. Kong, H. W. H. Hui, H. Peng, and W. W. B. Goh. Dealing with missing values in proteomics data. *PROTEOMICS*, page 2200092, 2022. ISSN 1615-9853. doi:10.1002/PMIC.202200092.

[Kong23] W. Kong, B. J. H. Wong, H. W. H. Hui, *et al.* ProJect: a powerful mixed-model missing value imputation method. *Briefings in Bioinformatics*, 24(4), 2023. ISSN 14774054. doi:10.1093/BIB/BBAD233.

[Korthauer19] K. Korthauer, P. K. Kimes, C. Duvallet, *et al.* A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20(1):1–21, 2019. ISSN 1474760X. doi:10.1186/S13059-019-1716-1.

[Koyuncu22] D. Koyuncu and B. Ulent Yener. Missing Value Knockoffs. *arXiv*, 2022. doi:10.48550/arxiv.2202.13054.

[Krämer86] W. Krämer and H. Sonnberger. *The linear regression model under test*. Springer Science & Business Media, 1986.

[Kurz22] M. S. Kurz. Vine copula based knockoff generation for high-dimensional controlled variable selection. *arXiv*, 2022. doi:10.48550/arXiv.2210.11196.

[Lassailly11] G. Lassailly, R. Caiazzo, A. Hollebecque, *et al.* Validation of noninvasive biomarkers (FibroTest, SteatoTest, and NashTest) for prediction of liver injury in patients with morbid obesity. *European Journal of Gastroenterology and Hepatology*, 23(6):499–506, 2011. ISSN 0954691X. doi:10.1097/MEG.0B013E3283464111.

[Lay06] J. O. Lay, R. Liyanage, S. Borgmann, and C. L. Wilkins. Problems with the "omics". *TrAC Trends in Analytical Chemistry*, 25(11):1046–1056, 2006. ISSN 0165-9936. doi:10.1016/J.TRAC.2006.10.007.

[Lazar15] C. Lazar. imputelcmd: a collection of methods for left-censored missing data imputation. *R package, version*, 2, 2015.

[Lazar16] C. Lazar, L. Gatto, M. Ferro, *et al.* Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *Journal of Proteome Research*, 15(4):1116–1125, 2016. ISSN 15353907. doi:10.1021/acs.jproteome.5b00981.

[Leppäaho17] E. Leppäaho, S. Kaski, and M. E. Khan. GFA: Exploratory Analysis of Multiple Data Sources with Group Factor Analysis Muhammad Ammad-ud-din. *Journal of Machine Learning Research*, 18:1–5, 2017. doi:10.5555/3122009.3122048.

[Levitsky17] L. I. Levitsky, M. V. Ivanov, A. A. Lobas, and M. V. Gorshkov. Unbiased False Discovery Rate Estimation for Shotgun Proteomics Based on the Target-Decoy Approach. *Journal of Proteome Research*, 16(2):393–397, 2017. ISSN 15353907. doi:10.1021/acs.jproteome.6b00144.

[Li20] Q. Li, K. Fisher, W. Meng, *et al.* GMSimpute: a generalized two-step Lasso approach to impute missing values in label-free mass spectrum analysis. *Bioinformatics*, 36(1):257–263, 2020. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTZ488.

[Li23] M. Li and G. K. Smyth. Neither random nor censored: estimating intensity-dependent probabilities for missing values in label-free proteomics. *Bioinformatics*, 39(5), 2023. ISSN 1367-4811. doi:10.1093/bioinformatics/btad200.

[Lim19] M. Y. Lim, J. A. Paulo, and S. P. Gygi. Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *Journal of Proteome Research*, 18(11):4020–4026, 2019. ISSN 15353907. doi:10.1021/ACS.JPROTEOME.9B00492.

[Lin21] A. Lin, D. L. Plubell, U. Keich, and W. S. Noble. Accurately Assigning Peptides to Spectra When only a Subset of Peptides Are Relevant. *Journal of Proteome Research*, 20(8):4153–4164, 2021. ISSN 15353907. doi:10.1021/acs.jproteome.1c00483.

[Lin22] A. Lin, T. Short, W. S. Noble, and U. Keich. Detecting more peptides from bottom-up mass spectrometry data via peptide-level target-decoy competition. *bioRxiv*, page 2022.05.11.491571, 2022. doi:10.1101/2022.05.11.491571.

[Linderman22] G. C. Linderman, J. Zhao, M. Roulis, *et al.* Zero-preserving imputation of single-cell RNA-seq data. *Nature Communications 2022 13:1*, 13(1):1–11, 2022. ISSN 2041-1723. doi:10.1038/s41467-021-27729-z.

[Little19] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019. doi:10.1002/9781119482260.

[Liu89] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, 1989. ISSN 00255610. doi:10.1007/BF01589116/METRICS.

[Liu16] Y. Liu, A. Beyer, and R. Aebersold. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, 165(3):535–550, 2016. ISSN 00928674. doi:10.1016/j.cell.2016.03.014.

[Liu21] M. Liu and A. Dongre. Proper imputation of missing values in proteomics datasets for differential expression analysis. *Briefings in Bioinformatics*, 22(3):1–17, 2021. ISSN 1467-5463. doi:10.1093/bib/bbaa112.

[Livera15] A. M. D. Livera, M. Sysi-Aho, L. Jacob, *et al.* Statistical Methods for Handling Unwanted Variation in Metabolomics Data. *Analytical Chemistry*, 87(7):3606–3615, 2015. ISSN 0003-2700. doi:10.1021/ac502439y.

[Love14] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, 2014. ISSN 1474760X. doi:10.1186/S13059-014-0550-8.

[Lowe17] R. Lowe, N. Shirley, M. Bleackley, *et al.* Transcriptomics technologies. *PLoS computational biology*, 13(5), 2017. ISSN 1553-7358. doi:10.1371/JOURNAL.PCBI.1005457.

[Luo09] R. Luo, C. M. Colangelo, W. C. Sessa, and H. Zhao. Bayesian Analysis of iTRAQ Data with Nonrandom Missingness: Identification of Differentially Expressed Proteins. *Statistics in Biosciences*, 1(2):228–245, 2009. ISSN 1867-1764. doi:10.1007/S12561-009-9013-2/METRICS.

[Ma17] C. Ma, S. Xu, G. Liu, *et al.* Improvement of peptide identification with considering the abundance of mRNA and peptide. *BMC Bioinformatics*, 18(1):1–8, 2017. ISSN 14712105. doi:10.1186/S12859-017-1491-5.

[Ma20] W. Ma, S. Kim, S. Chowdhury, *et al.* DreamAI: algorithm for the imputation of proteomics data. *bioRxiv*, page 2020.07.21.214205, 2020. doi:10.1101/2020.07.21.214205.

[Madej22] D. Madej, L. Wu, and H. Lam. Common Decoy Distributions Simplify False Discovery Rate Estimation in Shotgun Proteomics. *Journal of Proteome Research*, 21(2):339–348, 2022. ISSN 1535-3893. doi:10.1021/acs.jproteome.1c00600.

[Marasco22] L. E. Marasco and A. R. Kornblihtt. The physiology of alternative splicing. *Nature Reviews Molecular Cell Biology 2022 24:4*, 24(4):242–254, 2022. ISSN 1471-0080. doi:10.1038/s41580-022-00545-z.

[Martínez-Bartolomé08] S. Martínez-Bartolomé, P. Navarro, F. Martín-Maroto, *et al.* Properties of Average Score Distributions of SEQUEST: The Probability Ratio Method. *Molecular & Cellular Proteomics*, 7(6):1135–1145, 2008. ISSN 1535-9476. doi:10.1074/MCP.M700239-MCP200.

[Masselon03] C. Masselon, L. Paša-Tolić, S. W. Lee, *et al.* Identification of tryptic peptides from large databases using multiplexed tandem mass spectrometry: simulations and experimental results. *PROTEOMICS*, 3(7):1279–1286, 2003. ISSN 1615-9861. doi:10.1002/PMIC.200300448.

[McCullagh19] P. McCullagh and J. Nelder. *Generalized Linear Models*. Routledge, 2019. ISBN 9780203753736. doi:10.1201/9780203753736.

[Meng19] C. Meng, A. Basunia, B. Peters, *et al.* MOGSA: Integrative single sample gene-set analysis of multiple omics data. *Molecular and Cellular Proteomics*, 18(8):S153–S168, 2019. ISSN 15359484. doi:10.1074/mcp.TIR118.001251.

[Miao16] W. Miao, P. Ding, and Z. Geng. Identifiability of Normal and Normal Mixture Models with Nonignorable Missing Data. *Journal of the American Statistical Association*, 111(516):1673–1683, 2016. ISSN 1537274X. doi:10.1080/01621459.2015.1105808.

[Miller22] R. M. Miller, B. T. Jordan, M. M. Mehlferber, *et al.* Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biology*, 23(1):1–28, 2022. ISSN 1474760X. doi:10.1186/S13059-022-02624-Y.

[Moore02] R. E. Moore, M. K. Young, and T. D. Lee. Qscore: An algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, 13(4):378–386, 2002. ISSN 10440305. doi:10.1016/S1044-0305(02)00352-5.

[Moosa20] J. M. Moosa, S. Guan, M. F. Moran, and B. Ma. Repeat-Preserving Decoy Database for False Discovery Rate Estimation in Peptide Identification. *Journal of Proteome Research*, 19(3):1029–1036, 2020. ISSN 15353907. doi:10.1021/acs.jproteome.9b00555.

[Munteanu16] M. Munteanu, D. Tiniakos, Q. Anstee, *et al.* Diagnostic performance of FibroTest, SteatoTest and ActiTest in patients with NAFLD using the SAF score as histological reference. *Alimentary Pharmacology & Therapeutics*, 44(8):877–889, 2016. ISSN 1365-2036. doi:10.1111/APT.13770.

[Munteanu18] M. Munteanu, R. Pais, V. Peta, *et al.* Long-term prognostic value of the FibroTest in patients with non-alcoholic fatty liver disease, compared to chronic hepatitis C, B, and alcoholic liver disease. *Alimentary Pharmacology & Therapeutics*, 48(10):1117–1127, 2018. ISSN 1365-2036. doi:10.1111/APT.14990.

[Muth18] T. Muth and B. Y. Renard. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Briefings in Bioinformatics*, 19(5):954–970, 2018. ISSN 1467-5463. doi:10.1093/BIB/BBX033.

[Nguyen20] T.-b. Nguyen, J.-a. Chevalier, B. Thirion, *et al.* Aggregation of Multiple Knockoffs. *Proceedings of the 37th International Conference on Machine Learning*, 119:7283–7293, 2020.

[Nie06] L. Nie, G. Wu, F. J. Brockman, and W. Zhang. Integrated analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: zero-inflated Poisson regression models to predict abundance of undetected proteins. *Bioinformatics*, 22(13):1641–1647, 2006. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTL134.

[Niessen06] W. M. Niessen. Liquid Chromatography-Mass Spectrometry. *Liquid Chromatography-Mass Spectrometry*, 2006. doi:10.1201/9781420014549.

[Niu22] L. Niu, M. Thiele, P. E. Geyer, *et al.* Noninvasive proteomic biomarkers for alcohol-related liver disease. *Nature Medicine 2022 28:6*, 28(6):1277–1287, 2022. ISSN 1546-170X. doi:10.1038/s41591-022-01850-y.

[Oba03] S. Oba, M. A. Sato, I. Takemasa, *et al.* A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTG287.

[O'brien18] J. J. O'brien, H. P. Gunawardena, J. A. Paulo, *et al.* The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *The annals of applied statistics*, 12(4):2075, 2018. ISSN 19417330. doi:10.1214/18-AOAS1144.

[Ochoteco Asensio22] J. Ochoteco Asensio, M. Verheijen, and F. Caiment. Predicting missing proteomics values using machine learning: Filling the gap using transcriptomics and other biological features. *Computational and Structural Biotechnology Journal*, 20:2057–2069, 2022. ISSN 2001-0370. doi:10.1016/J.CSBJ.2022.04.017.

[O'Leary16] N. A. O'Leary, M. W. Wright, J. R. Brister, *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2016. ISSN 0305-1048. doi:10.1093/NAR/GKV1189.

[Pappireddi19] N. Pappireddi, L. Martin, and M. Wühr. A Review on Quantitative Multiplexed Proteomics. *Chembiochem : a European journal of chemical biology*, 20(10):1210, 2019. ISSN 14397633. doi:10.1002/CBIC.201800650.

[Peng03] J. Peng, J. E. Elias, C. C. Thoreen, *et al.* Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome. *Journal of Proteome Research*, 2(1):43–50, 2003. ISSN 1535-3893. doi:10.1021/pr025556v.

[Perkins99] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999. ISSN 0173-0835. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

[Pinheiro96] J. C. Pinheiro and D. M. Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing 1996*, 6(3):289–296, 1996. ISSN 1573-1375. doi:10.1007/BF00140873.

[Pounds06] S. Pounds and C. Cheng. Robust estimation of the false discovery rate. *Bioinformatics*, 22(16):1979–1987, 2006. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTL328.

[Poynard05] T. Poynard, F. Imbert-Bismut, M. Munteanu, and V. Ratziu. FibroTest-FibroSURE™: towards a universal biomarker of liver fibrosis? *Expert Review of Molecular Diagnostics*, 5(1):15–21, 2005. ISSN 14737159. doi:10.1586/14737159.5.1.15.

[Prieto20] G. Prieto and J. Vázquez. Protein Probability Model for High-Throughput Protein Identification by Mass Spectrometry-Based Proteomics. *Journal of Proteome Research*, 19(3):1285–1297, 2020. ISSN 1535-3893. doi:10.1021/acs.jproteome.9b00819.

[R Core Team13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[Ramus16] C. Ramus, A. Hovasse, M. Marcellin, *et al.* Benchmarking quantitative label-free LC–MS data processing workflows using a complex spiked proteomic standard dataset. *Journal of Proteomics*, 132:51–62, 2016. ISSN 1874-3919. doi:10.1016/J.JPROT.2015.11.011.

[Ritchie15] M. E. Ritchie, B. Phipson, D. Wu, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015. ISSN 0305-1048. doi:10.1093/NAR/GKV007.

[Robert99] C. P. Robert and G. Casella. *Monte Carlo Integration*, pages 71–138. Springer New York, New York, NY, 1999. ISBN 978-1-4757-3071-5. doi:10.1007/978-1-4757-3071-5_3.

[Rohart17] F. Rohart, B. Gautier, A. Singh, and K.-A. L. Cao. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11):e1005752, 2017. ISSN 1553-7358. doi:10.1371/JOURNAL.PCBI.1005752.

[Romano06] J. P. Romano, A. M. Shaikh, *et al.* On stepdown control of the false discovery proportion. In *Optimality*, pages 33–50. Institute of Mathematical Statistics, 2006.

[Romano19] Y. Romano, M. Sesia, and E. Candès. Deep Knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2019. ISSN 1537274X. doi:10.1080/01621459.2019.1660174.

[Ropers21] D. Ropers, Y. Couté, L. Faure, *et al.* Multiomics Study of Bacterial Growth Arrest in a Synthetic Biology Application. *ACS Synthetic Biology*, 10(11):2910–2926, 2021. ISSN 21615063. doi:10.1021/ACSSYNBIO.1C00115.

[Ryu14] S. Y. Ryu, W. J. Qian, D. G. Camp, *et al.* Detecting differential protein expression in large-scale population proteomics. *Bioinformatics*, 30(19):2741–2746, 2014. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTU341.

[Sesia19] M. Sesia, C. Sabatti, and E. J. Candès. Gene hunting with hidden Markov model knockoffs. *Biometrika*, 106(1):1–18, 2019. ISSN 0006-3444. doi:10.1093/BIOMET/ASY033.

[Shah17] J. S. Shah, S. N. Rai, A. P. DeFilippis, *et al.* Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. *BMC Bioinformatics*, 18(1):1–13, 2017. ISSN 14712105. doi:10.1186/S12859-017-1547-6/TABLES/4.

[Shanmugam14] A. K. Shanmugam, A. K. Yocum, and A. I. Nesvizhskii. Utility of RNA-seq and GPMDB protein observation frequency for improving the sensitivity of protein identification by tandem MS. *Journal of Proteome Research*, 13(9):4113–4119, 2014. ISSN 15353907. doi:10.1021/PR500496P.

[Shen22] M. Shen, Y. T. Chang, C. T. Wu, *et al.* Comparative assessment and novel strategy on methods for imputing proteomics data. *Scientific Reports 2022 12:1*, 12(1):1–11, 2022. ISSN 2045-2322. doi:10.1038/s41598-022-04938-0.

[Shi21] Z. Shi, B. Wen, Q. Gao, and B. Zhang. Feature Selection Methods for Protein Biomarker Discovery from Proteomics or Multi-omics Data. *Molecular & Cellular Proteomics*, page 100083, 2021. ISSN 15359476. doi:10.1016/j.mcpro.2021.100083.

[Silverman20] J. D. Silverman, K. Roche, S. Mukherjee, and L. A. David. Naught all zeros in sequence count data are the same. *Computational and Structural Biotechnology Journal*, 18:2789, 2020. ISSN 20010370. doi:10.1016/J.CSBJ.2020.09.014.

[Song20] M. Song, J. Greenbaum, J. Luttrell, *et al.* A Review of Integrative Imputation for Multi-Omics Datasets. *Frontiers in Genetics*, 11(October):1–15, 2020. ISSN 16648021. doi:10.3389/fgene.2020.570255.

[Sportisse20] A. Sportisse, C. Boyer, and J. Josse. Estimation and Imputation in Probabilistic Principal Component Analysis with Missing Not At Random Data. *Advances in Neural Information Processing Systems*, 33:7067–7077, 2020. doi:10.5555/3495724.3496317.

[Stahl96] D. C. Stahl, K. M. Swiderek, M. T. Davis, and T. D. Lee. Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *Journal of the American Society for Mass Spectrometry*, 7(6):532–540, 1996. ISSN 10440305. doi:10.1016/1044-0305(96)00057-8.

[Stekhoven12] D. J. Stekhoven and P. Bühlmann. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012. ISSN 13674803. doi:10.1093/BIOINFORMATICS/BTR597.

[Stephens17] M. Stephens. False discovery rates: A new deal. *Biostatistics*, 18(2):275–294, 2017. ISSN 14684357. doi:10.1093/biostatistics/kxw041.

[Storey02] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002. ISSN 1467-9868. doi:10.1111/1467-9868.00346.

[Storey03] J. D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003. ISSN 0090-5364. doi:10.1214/AOS/1074290335.

[Storey04] J. D. Storey, J. E. Taylor, and D. Siegmund. Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1):187–205, 2004. ISSN 1369-7412. doi:10.1111/J.1467-9868.2004.00439.X.

[Stuart19] T. Stuart, A. Butler, P. Hoffman, *et al.* Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21, 2019. ISSN 10974172. doi:10.1016/j.cell.2019.05.031.

[Stuwe14] E. Stuwe, K. F. Tóth, and A. A. Aravin. Small but sturdy: small RNAs in cellular memory and epigenetics. *Genes & Development*, 28(5):423, 2014. ISSN 08909369. doi:10.1101/GAD.236414.113.

[Sun06] L. Sun, R. V. Craiu, A. D. Paterson, and S. B. Bull. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, 30(6):519–530, 2006. ISSN 1098-2272. doi:10.1002/GEPI.20164.

[Tariq21] M. U. Tariq, M. Haseeb, M. Aledhari, *et al.* Methods for Proteogenomics Data Analysis, Challenges, and Scalability Bottlenecks: A Survey. *IEEE Access*, 9:5497–5516, 2021. doi:10.1109/ACCESS.2020.3047588.

[Taylor13] S. L. Taylor, G. S. Leiserowitz, and K. Kim. Accounting for Undetected Compounds in Statistical Analyses of Mass Spectrometry 'Omic Studies. *Statistical applications in genetics and molecular biology*, 12(6):703, 2013. doi:10.1515/SAGMB-2013-0021.

[Taylor22] S. Taylor, M. Ponzini, M. Wilson, and K. Kim. Comparison of imputation and imputation-free methods for statistical analysis of mass spectrometry data with missing data. *Briefings in Bioinformatics*, 23(1):1–11, 2022. ISSN 14774054. doi:10.1093/BIB/BBAB353.

[Thompson03] A. Thompson, J. Schäfer, K. Kuhn, *et al.* Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, 75(8):1895–1904, 2003. ISSN 00032700. doi:10.1021/AC0262560.

[Tibshirani96] R. Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. ISSN 2517-6161. doi:10.1111/J.2517-6161.1996.TB02080.X.

[Torres-García09] W. Torres-García, W. Zhang, G. C. Runger, *et al.* Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins. *Bioinformatics*, 25(15):1905–1914, 2009. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTP325.

[Torres-García11] W. Torres-García, S. D. Brown, R. H. Johnson, *et al.* Integrative analysis of transcriptomic and proteomic data of Shewanella oneidensis : missing value imputation using temporal datasets. *Molecular BioSystems*, 7(4):1093–1104, 2011. doi:10.1039/C0MB00260G.

[Troyanskaya01] O. Troyanskaya, M. Cantor, G. Sherlock, *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/17.6.520.

[Tsiatsiani15] L. Tsiatsiani, A. J. R. Heck, A. J. R. Heck, and P. Bijvoet. Proteomics beyond trypsin. *The FEBS Journal*, 282(14):2612–2626, 2015. ISSN 1742-4658. doi:10.1111/FEBS.13287.

[Tusher01] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001. ISSN 0027-8424. doi:10.1073/pnas.091062498.

[Tyanova16] S. Tyanova, T. Temu, and J. Cox. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols 2016 11:12*, 11(12):2301–2319, 2016. ISSN 1750-2799. doi:10.1038/nprot.2016.136.

[Uversky13] V. N. Uversky. Posttranslational Modification. *Brenner's Encyclopedia of Genetics: Second Edition*, pages 425–430, 2013. doi:10.1016/B978-0-12-374984-0.01203-1.

[van Buuren11] S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. doi:10.18637/jss.v045.i03.

[Venter01] J. C. Venter, M. D. Adams, E. W. Myers, *et al.* The Sequence of the Human Genome. *Science*, 291(5507):1304–1351, 2001. ISSN 0036-8075. doi:10.1126/science.1058040.

[Verboven07] S. Verboven, K. V. Branden, and P. Goos. Sequential imputation for missing values. *Computational Biology and Chemistry*, 31(5-6):320–327, 2007. ISSN 1476-9271. doi:10.1016/J.COMPBIOLCHEM.2007.07.001.

[Verheggen16] K. Verheggen, L. Martens, F. S. Berven, *et al.* Database Search Engines: Paradigms, Challenges and Solutions. *Advances in experimental medicine and biology*, 919:147–156, 2016. ISSN 0065-2598. doi:10.1007/978-3-319-41448-5_6.

[Vidova17] V. Vidova and Z. Spacil. A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. *Analytica Chimica Acta*, 964:7–23, 2017. ISSN 0003-2670. doi:10.1016/J.ACA.2017.01.059.

[Vilallongue22] N. Vilallongue, J. Schaeffer, A. M. Hesse, *et al.* Guidance landscapes unveiled by quantitative proteomics to control reinnervation in adult visual system. *Nature Communications*, 13(1):1–20, 2022. ISSN 20411723. doi:10.1038/s41467-022-33799-4.

[Villate21] A. Villate, M. San Nicolas, M. Gallastegi, *et al.* Review: Metabolomics as a prediction tool for plants performance under environmental stress. *Plant Science*, 303:110789, 2021. ISSN 0168-9452. doi:10.1016/J.PLANTSCI.2020.110789.

[Wang20] S. Wang, W. Li, L. Hu, *et al.* NAguideR: Performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. *Nucleic Acids Research*, 48(14):e83, 2020. ISSN 13624962. doi:10.1093/nar/gkaa498.

[Wang22] J. Wang, X. Gong, M. Hu, and L. Zhao. Improved GSimp: A Flexible Missing Value Imputation Method to Support Regulatory Bioequivalence Assessment. *Annals of Biomedical Engineering 2022*, pages 1–11, 2022. ISSN 1573-9686. doi:10.1007/S10439-022-03070-4.

[Webb-Robertson15] B. J. M. Webb-Robertson, H. K. Wiberg, M. M. Matzke, *et al.* Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of Proteome Research*, 14(5):1993–2001, 2015. ISSN 15353907. doi:10.1021/pr501138h.

[Webel23] H. Webel, L. Niu, A. B. Nielsen, *et al.* Imputation of label-free quantitative mass spectrometry-based proteomics data using self supervised deep learning. *bioRxiv*, page 2023.01.12.523792, 2023. doi:10.1101/2023.01.12.523792.

[Wei18] R. Wei, J. Wang, E. Jia, *et al.* GSimp: A Gibbs sampler based left-censored missing value imputation approach for metabolomics studies. *PLoS Computational Biology*, 14(1):e1005973, 2018. ISSN 15537358. doi:10.1371/journal.pcbi.1005973.

[Wieczorek17] S. Wieczorek, F. Combes, C. Lazar, *et al.* DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics*, 33(1):135–136, 2017. ISSN 1367-4803. doi:10.1093/BIOINFORMATICS/BTW580.

[Wieczorek19] S. Wieczorek, F. Combes, H. Borges, and T. Burger. Protein-level statistical analysis of quantitative label-free proteomics data with ProStaR. *Methods in Molecular Biology*, 1959:225–246, 2019. ISSN 10643745. doi:10.1007/978-1-4939-9164-8_15.

[Wikipedia23] Wikipedia. `https://en.wikipedia.org/wiki/List_of_omics_topics_in_biology#cite_note-Omics.org-2`, 2023.

[Xia23] L. Xia, C. Lee, and J. J. Li. scDEED: a statistical method for detecting dubious 2D single-cell embeddings. *bioRxiv*, page 2023.04.21.537839, 2023. doi:10.1101/2023.04.21.537839.

[Xie11] F. Xie, T. Liu, W.-J. Qian, *et al.* Liquid Chromatography-Mass Spectrometry-based Quantitative Proteomics. *Journal of Biological Chemistry*, 286(29):25443–25449, 2011. ISSN 0021-9258. doi:10.1074/JBC.R110.199703.

[Xing21] X. Xing, Z. Zhao, and J. S. Liu. Controlling False Discovery Rate Using Gaussian Mirrors. *Journal of the American Statistical Association*, 0(0):1–20, 2021. ISSN 0162-1459. doi:10.1080/01621459.2021.1923510.

[Yekutieli99] D. Yekutieli and Y. Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1-2):171–196, 1999. ISSN 0378-3758. doi:10.1016/S0378-3758(99)00041-5.

[Zahn-Zabal20] M. Zahn-Zabal, P. A. Michel, A. Gateau, *et al.* The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Research*, 48(D1):D328–D334, 2020. ISSN 0305-1048. doi:10.1093/NAR/GKZ995.

[Zeileis15] A. Zeileis and T. Hothorn. Diagnostic Checking in Regression Relationships. *R News*, 2(3):7–10, 2015.

[Zhang15] Y. Zhang, Z. Wen, M. P. Washburn, and L. Florens. Improving Label-Free Quantitative Proteomics Strategies by Distributing Shared Peptides and Stabilizing Variance. *Analytical Chemistry*, 87(9):4749–4756, 2015. doi:10.1021/AC504740P.

[Zhou20] Z. Zhou, C. Ye, J. Wang, and N. R. Zhang. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nature Communications*, 11(1):1–10, 2020. ISSN 20411723. doi:10.1038/s41467-020-14391-0.

[Zou05] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. ISSN 1467-9868. doi:10.1111/J.1467-9868.2005.00503.X.

List of Figures

List of Tables

Résumé

Si le monde du vivant était un jeu de LEGO, les protéines en seraient les briques de base. Ainsi, être capable d'identifier et quantifier toutes ces protéines à l'échelle de quelques cellules, ou d'un être vivant (c'est-à-dire leur "protéome"), permet aux biologistes de mieux comprendre leur physiologie à un temps donné. Pour ce faire, on utilise un spectromètre de masse, qui produit de nombreuses données complexes et sujettes à de nombreux biais. Des algorithmes doivent alors automatiquement en extraire de l'information utile et interprétable par les biologistes. Le but de ma thèse est d'améliorer de tels algorithmes, tout en faisant en sorte qu'ils soient statistiquement robustes. J'ai contribué d'une part à limiter le taux de fausses découvertes dans les expériences de protéomiques; et d'autre part, j'ai développé un algorithme permettant d'inférer des valeurs d'abondance de protéines manquantes dans des tableaux quantitatifs résultant de ces expériences.

Mots-clés : Biostatistiques, Protéomique, Imputation de Valeurs manquantes, Contrôle du FDR, Spectrométrie de masse, Transcriptomique

Abstract

Proteins are the basic molecular building blocks of life. Thus, being able to identify and quantify all these proteins at the scale of a few cells, or of an organism, (i.e., their "proteome"), allows biologists to better understand their physiology at a given time. To do this, a mass spectrometer is involved: it produces many complex data that are subject to numerous biases. Algorithms must then automatically extract useful information for subsequent biological interpretation. The goal of my thesis is to develop these algorithms, while ensuring they are statistically robust. On the one hand, I have contributed to controlling for the false discovery rate in proteomics experiments, and on the other hand, I have developed an algorithm to infer the missing abundance values of proteins in quantitative tables resulting from these experiments.

Keywords : Biostatistics, Proteomics, Missing Value Imputation, FDR control, Mass Spectrometry, Transcriptomics